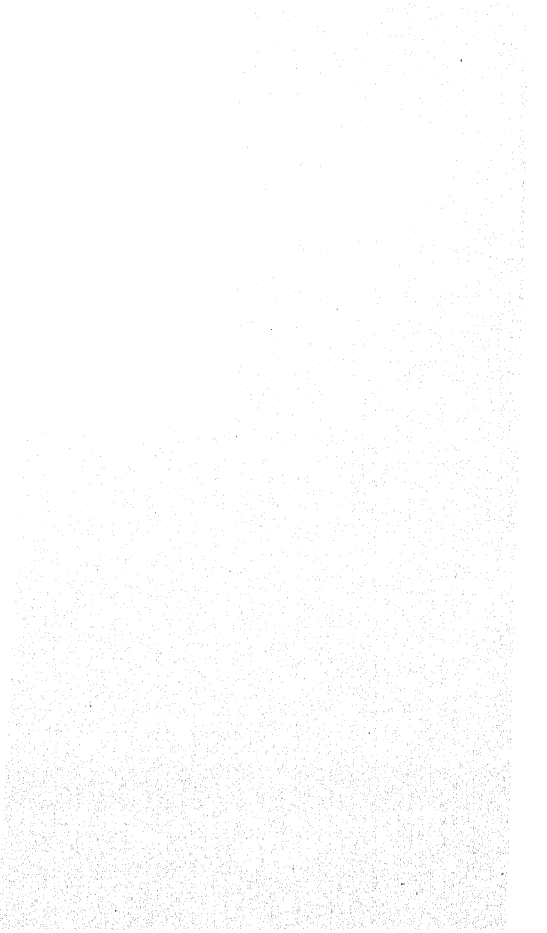
PERSONNEL RESEARCH AND TEST DEVELOPMENT



PERSONNEL RESEARCH AND TEST DEVELOPMENT IN THE BUREAU OF NAVAL PERSONNEL

BY THE
STAFF, TEST AND RESEARCH SECTION
IN COOPERATION WITH
N.D.R.C. PROJECT N-106 AND THE
COLLEGE ENTRANCE EXAMINATION BOARD

EDITED BY
DEWEY B. STUIT
LT. COMDR., U.S.N.R.
ASSISTANT OFFICER-IN-CHARGE
TEST AND RESEARCH SECTION

PRINCETON, NEW JERSEY
PRINCETON UNIVERSITY PRESS
1947

COPYRIGHT, 1947, BY PRINCETON UNIVERSITY PRESS
LONDON: GEOFFREY CUMBERLEGE, OXFORD UNIVERSITY PRESS

PRINTED IN THE UNITED STATES OF AMERICA
BY AMERICAN BOOK–STRATFORD PRESS, INC., NEW YORK

The opinions expressed in this book are those of the authors and are not to be construed as reflecting the policies or opinions of the Navy Department.



FOREWORD

THE Navy that fought in World War II was the most technical and highly organized of any navy in history. The combat success of a fleet, task force, or unit depended in large part on the individual skill of the officers and men manning battle stations and on their ability to work as a team. This success, in the many complex operations, was in no small measure due to the fact that personnel had been placed in jobs for which they had aptitude and for which they had been trained. Modern selection, classification, and training methods were an essential factor in solving the Navy's tremendous manpower problem in World War II.

The tests and research studies reported in this book are the major personnel research activities conducted by the Bureau of Naval Personnel with the assistance of the National Defense Research Committee Project N-106 and the College Entrance Examination Board. The writers of the chapters have not only described the results of past work but have called attention to some of the many personnel problems still confronting the Navy. The findings presented in this publication are valuable not only for the immediately useful data they provide but even more for their stimulation of future research.

Many of the officers and civilian personnel who have contributed to this work have returned to their peacetime pursuits. The Bureau of Naval Personnel appreciates their contribution in increasing the efficiency of its program and in laying the ground work for study, research, and development in the years that lie ahead.

W. G. COOPER
Captain, U.S.N.
Director
Research Activity

PREFACE

THIS book is intended primarily for personnel psychologists in military organizations, educational institutions, public personnel administration, business, and industry. While few distinctly new techniques are presented, there is given an account of the experience of one military personnel research organization with the application of established techniques to the varied problems of military selection, classification, and training. This experience should be of value to students of personnel psychology and to individuals responsible for planning personnal research programs.

The purpose of this book is to present an evaluative summary of the personnel research and test development completed during World War II by the Test and Research Section of the Bureau of Naval Personnel in cooperation with the National Defense Research Committee Project N-106 and the College Entrance Examination Board. Since the Test and Research Section and the latter organizations worked cooperatively on many projects, there is no attempt made here to identify the specific accomplishments of each.

Part I presents a history of personnel research and test development in the Bureau of Naval Personnel and a brief account of the Bureau's selection, classification and training programs. These chapters are intended to give the reader the necessary background for the chapters which follow; by first reading these, the reader will be in a better position to see how tests, technical personnel aids, and the results of research were put to administrative use. As pointed out in Chapter I, the Test and Research Section is a research organization and as such does not have administrative control over selection, classification, or training programs.

Part II describes the construction of the selection and classification instruments developed for use by the Bureau of Naval Personnel. The Basic Test Battery administered to over 2,000,000 recruits and the Officer Qualification and Officer Classification Tests administered to over 100,000 officers are described in this part. The research on the construction and validation of measures of personal adjustment is also included in this part.

Part III is concerned with the prediction of success in Navy training programs. The selection of individuals for various types of schools was a major personnel problem both on the officer and enlisted level. While it would have been highly desirable to correlate test scores and other qualifications data of individuals with success on the job, most of the Bureau's validation studies used as criteria

of success grades received in school. On the basis of these prediction studies, selection requirements were formulated or revised.

Part IV describes the development and use of achievement tests in Navy training programs. The emphasis in this part is on the effect that standardized achievement tests have had on the quality of Navy instruction and the administration of schools. No attempt is made to give a description of all the achievement tests which were developed or of the studies made involving the use of achievement tests.

Follow-up studies and plans for future research are included in Part V. Since the nature of the criterion is one of the most important aspects of a follow-up study, the first chapter is devoted to a discussion of this subject. The data presented in Chapter XX are unique in that they constitute the results of the first systematic study undertaken by the Navy on the relationship between the background factors of enlisted personnel and performance aboard ship. Chapter XXI describes the first systematic studies sponsored by the Navy to obtain reactions of personnel to the Navy's classification and training programs. The purpose of the final chapter is to discuss briefly some of the classification and training problems which still remain in the Navy.

The terms qualification, selection, and classification are used frequently in this book. During the early part of the war the term qualification was generally applied to determining whether an individual was suitable for the naval service; selection was applied to picking candidates for school training; determining qualifications for duty was termed classification. With the passage of time, the process of determining qualifications for both training and billet assignments was termed classification and there was an inclination to use the terms qualification and selection interchangeably. Because of their varied use, there is no attempt made in this book to attach precise meaning to the terms, although the preference of the authors is to apply the term *selection* to determining qualification for the naval service and *classification* to determining qualifications for training and billet assignments.

The materials presented in this publication do not constitute a complete account of the work of either the Test and Research Section or of the NDRC Project N-106 and the College Entrance Examination Board; but there is given in the appendix a complete list of the tests produced or officially approved by the Test and Research Section, a list of research studies completed by the Section, and those reports of NDRC Project N-106 which constituted source material for research discussed in this volume. Experimental tests and informal research reports have been omitted from these lists. Tests and research data obtained from NDRC projects other than

N-106 are identified as having been "furnished by projects of the National Defense Research Committee".

Special mention should be made of the fact that this book does not contain accounts of all the personnel research done in or for the Navy. Research in aviation psychology and clinical psychology was under the cognizance of the Bureau of Medicine and Surgery, as was a major share of the work on the selection, classification, and training of submarine personnel. Research on the design and use of aviation training devices was under the direction of the Special Devices Division, Bureau of Aeronautics. Research, in addition to that sponsored by the Bureau of Naval Personnel, Bureau of Medicine and Surgery, or Special Devices Division, was done by personnel stationed in the Navy Department in Washington and at various field activities and on units afloat. Finally, there is the extensive contribution of the many projects of the National Defense Research Committee, not only to the Bureau of Naval Personnel's program but to the many field stations and training activities.

The studies and tests discussed in this volume are the result of the team effort of a large number of individuals. In a sense the authors of the various chapters are serving as reporters or interpreters of the work done in particular areas. Since it would be extremely difficult to make appropriate acknowledgements (except in a few instances) for the specific contributions made by each person, a group acknowledgement is made by the listing of names in Appendix A.

The success of the Bureau's test construction and research program is in large part due to the capable leadership of Lt. Comdr. Ray N. Faulkner, Officer-in-Charge of the Test and Research Section. His ability to weld the work of a large number of individuals into a team effort, to plan the program of the Section and to coordinate its work with that of other units of the Bureau was outstanding. His criticisms and suggestions in the preparation of this manuscript have been very helpful.

The accomplishments in research and test construction would not have been possible without the assistance and cooperation of various persons in the Training, Officer Personnel, Enlisted Personnel, and Research Activities of the Bureau of Naval Personnel. Grateful acknowledgement is made to Commander Alvin C. Eurich, Captain Carleton R. Adams, Captain Lot Ensey, successive directors of the Standards and Curriculum Division, Training, and the successive assistant directors, Lt. Comdr. Frank H. Bowles, Commander Paul A. Jones and Captain A. John Bartky for their assistance in planning and carrying out the program of the Test and Research Section; to Captain James B. Hogle and Lt. Comdr. James A. McCain of the

Enlisted Classification Section, and Lt. Comdr. John H. Cornehlsen and Lt. Comdr. David G. Price of the Officer Selection Unit for their cooperation in providing the administrative facilities for experimental test tryouts and in making available some of the basic data used in research studies; finally to Captain William G. Cooper, Director, Research Activity, under whose leadership the Test and Research Section was elevated to the division level and the first phases of the peacetime test development and personnel research program were planned.

Special thanks are also due Mr. John Stalnaker, Mr. Henry Chauncey, and Dr. Harold O. Gulliksen of the National Defense Research Committee Project N-106 and the College Entrance Examination Board, who gave outstanding technical assistance in carrying through the Bureau's test construction and personnel research program; to Dr. Charles W. Bray and Dr. Dael L. Wolfle of the Applied Psychology Panel for their interest and encouragement in the preparation of this book and for assistance in making available to the Bureau the research findings and instruments developed by the projects of the National Defense Research Committee.

In the editorial preparation of this volume the virtual collaboration of Dr. Helen R. Haggerty was of the greatest assistance. She went over the manuscript in great detail and was completely responsible for organizing the tables in consistent format. In addition she offered many constructive suggestions with respect to the organization and presentation of the material.

DEWEY B. STUIT
Editor

TABLE OF CONTENTS

PART I

THE NAVY'S SELECTION, CLASSIFICATION AND TRAINING PROGRAMS

CHAPTER I. Personnel Research and Test Development in the Bureau of Naval Personnel: History and Scope of the Program. 3
RAY N. FAULKNER AND HELEN R. HAGGERTY

CHAPTER II. Selection and Classification of Officer Personnel. 12
JOHN H. CORNEHLSEN

CHAPTER III. Selection and Classification of Enlisted Personnel. 21
CHARLES E. ODELL

CHAPTER IV. The Program for Training Officer Personnel. 31
EUGENE D. CARSTATER

CHAPTER V. The Program for Training Enlisted Personnel. 41
HOWARD T. BATCHELDER

PART II

THE CONSTRUCTION, STANDARDIZATION, AND USE OF SELECTION AND CLASSIFICATION TESTS

CHAPTER VI. Basic Tests for Enlisted Personnel. 53
GUY L. BOND AND JOSEPH MILLER

CHAPTER VII. Basic Tests for Officer Personnel. 84
JOSEPH MILLER AND WILLIAM A. OWENS

CHAPTER VIII. Special Aptitude Tests. 112
DEWEY B. STUIT AND DANIEL D. FEDER

CHAPTER IX. Measures of Personal Adjustment. 126
MILTON WEXLER

PART III

PREDICTION OF SUCCESS IN TRAINING

CHAPTER X. Prediction of Success in Primary Officer Training Schools. 177
HERBERT S. CONRAD AND GERALD V. LANNHOLM

CHAPTER XI. Prediction of Success in Advanced Officer Training Programs. 216
JAMES W. MAUCKER

CHAPTER XII. Prediction of Success in Elementary Schools for Enlisted Personnel. 233
ROYAL F. BLOOM AND EVERETT G. BRUNDAGE

CHAPTER XIII. Prediction of Success in Advanced Service Schools. 262
JAMES F. CURTIS

PART IV

THE CONSTRUCTION AND USE OF ACHIEVEMENT MEASURES

CHAPTER XIV. Services Provided to Navy Training through Achievement Examinations. 287
DAVID G. RYANS

CHAPTER XV. Achievement Examinations for Elementary Enlisted Schools. 295
RUTHERFORD B. PORTER AND CHARLES M. HARSH

CHAPTER XVI. Achievement Examinations for Officer Schools. 315
EUGENE D. CARSTATER

CHAPTER XVII. The Measurement of Achievement in the Radio Technician Training Program. 331
DANIEL D. FEDER AND WILLIAM R. LAWRENCE

CHAPTER XVIII. Advancement in Rating Examinations. 344
RUTH M. CRUIKSHANK AND WESLEY C. DARLING

PART V

FOLLOW-UP STUDIES OF TRAINING AND CLASSIFICATION TECHNIQUES

CHAPTER XIX. Problems in Establishing Criterion Measures. 357
HAROLD P. BECHTOLDT

CHAPTER XX. Prediction of Performance of Enlisted Personnel Aboard Ship. 380
HAROLD P. BECHTOLDT, JAMES W. MAUCKER, AND DEWEY B. STUIT

CHAPTER XXI. Information Surveys as Evaluative Devices. 410
C. ROBERT PACE

CHAPTER XXII. Problems for Further Study. 433
NORMAN FREDERIKSEN, EUGENE D. CARSTATER, AND DEWEY B. STUIT

APPENDICES

A. Lists of Personnel. 457
A-1. Staff and Organization of Test and Research Section, 15 August 1945. 457
A-2. Staff of NDRC Project N-106 and Research Department of the College Entrance Examination Board. 458
B. List of Selection and Classification Tests, Advancement Examinations, Achievement Examinations, and Technical Personnel Aids Developed or Approved for Official Use by the Test and Research Section. 459
C. Lists of Research Studies. 465
C-1. List of research projects completed by the Test and Research Section. 465

C-2. Topical list of researches by NDRC Project N-106, College Entrance Examination Board, on the Navy's Aptitude Testing Program. 468
D. Sample Test Materials. 470
D-1. Sample directions for administering a performance Test. 470
D-2. Sample directions for administering an identification Test. 476
E. Technical Appendix. 482
E-1. Selection of items for the Officer Qualification Test, Forms 1, 2, 3. 482
E-2. Directions for conversion of raw scores on Achievement Tests to grades in the Navy 0-99 point scale. 485
E-3. Construction and use of abacs, by Donald A. Peterson and Harold O. Gulliksen. 488
E-4. Table for estimating population correlation coefficient from correlation coefficient obtained on a restricted sample. 506
INDEX 509

TABLES

Table 1-v. Base Pay for Navy Rates....42
Table 1-vi. Mean scores on Forms 2 and 3 of the Basic Test Battery for five Naval Training Centers....67
Table 2-vi. Estimated reliability coefficients of Basic Test Battery based upon data obtained from routine administration to recruits....69
Table 3-vi. Reliability coefficients of Basic Test Battery (Fleet Edition) based on data from recruit performance....70
Table 4-vi. Correlation coefficients of scores on Basic Test Battery with age and with highest school grade completed....71
Table 5-vi. Intercorrelations among tests of the Basic Test Battery based on data from routine administration to recruits....73
Table 1-vii. Officer Qualification Test, SE 0-1 and Form 1. Means and standard deviations (raw scores) on the subtests and total test....89
Table 2-vii. Officer Qualification Test, SE 0-1 and Form 1. Reliability coefficients and intercorrelations of subtests and total test....89
Table 3-vii. Officer Qualification Test, Forms 1, 2 and 3. Means and standard deviations (raw scores) on the subtests and total test for men and women....91
Table 4-vii. Officer Qualification Test, Forms 2 and 3. Reliability coefficients of subtests and total test....91
Table 5-vii. Officer Qualification Test, Forms 2 and 3. Intercorrelations of subtests and total test in Indoctrination School (Men) and Reserve Midshipmen's school (WR)....93
Table 6-vii. Officer Qualification Test, Forms 2 and 3. Intercorrelations, means and standard deviations (raw scores) of subtests and total test, national samples....94
Table 7-vii. Officers' Selective Examination (SE O-1). Comparison of means and standard deviations (raw scores) for limited vs. unlimited time....96
Table 8-vii. Officer Qualification Test, Form O-2. Comparison of mean scores and standard deviations (raw scores) for samples of men (in Indoctrination School), and women (in a Reserve Midshipmen's School [WR])....98
Table 9-vii. Officer Qualification Test, Forms 1, 2, and 3. Correlation coefficients of scores with age in years of students in Indoctrination School and Reserve Midshipmen's School (WR)....100
Table 10-vii. Officer Classification Test, Form X-1. Means, standard deviations, intercorrelations, and reliability coefficients....104
Table 11-vii. Officer Classification Test. Correlation coefficients between test scores and course grades at the Officer Submarine School....106
Table 12-vii. College Qualifying Test (C-1). Means, standard deviations, reliability coefficients, and intercorrelations of sections and total test....110

Table 1-VIII. Part-whole and interpart correlation coefficients for the CIC Aptitude Test 115
Table 2-VIII. Reliability coefficients for the parts of and the total CIC Aptitude Test 115
Table 3-VIII. Means and standard deviations for the Pre-Radar Officer Aptitude Test Battery administered to general duty and specialist duty midshipmen 117
Table 4-VIII. Correlation coefficients between scores on the Basic Test Battery, Form 3, and the Radio Technician Selection Test, Form 9A, for an unselected sample of Regular Navy recruits..... 120
Table 5-VIII. Means and standard deviations for the Basic Test Battery, Form 3, for Navy inductees and Regular Navy recruits..... 121
Table 1-IX. Description of research projects in the validation of psychiatric screening tests..... 143
Table 2-IX. Enlisted Personal Inventory, Parts 1 and 2: percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Receiving Station A, Sample 1..... 148
Table 3-IX. Enlisted Personal Inventory, Form 2, Parts 1 and 2: percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Naval Training Center B, Groups Ib, IIb, IIIb.... 149
Table 4-IX. Billet Qualifications Blank, Form X-2(M), Scale N and Scale N plus Scale L: percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Naval Training Center B, Groups Ib, IIb, IIIb..... 154
Table 5-IX. Billet Qualifications Blank, Form X-2(M), Scale N and Scale N plus Scale L: percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Receiving Station C..... 158
Table 6-IX. Experience Comparison Index, Form X-1, and Billet Qualifications Blank, Form X-2(M), Scale N plus Scale L: percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Receiving Station C..... 163
Table 7-IX. Personal Check List, Form X-4, and Enlisted Personal Inventory, Form 2, Part 1: percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population tested. Receiving Station A, Samples 1 and 2..... 167
Table 1-X. Correlations between Forms 1, 2, and 3 of the Officer Qualification Test and components of the final average grade (Indoctrination School) 182
Table 2-X. Means and standard deviations of course-grades and of Officer-Aptitude Ratings (Indoctrination School)..... 182
Table 3-X. Median correlations between subtests of the Officer Qualification Test and school grades (Indoctrination School)..... 183

Table 4-x. Simple vs. multiple correlation between parts of Officer Qualification Test and school grades (Indoctrination School).....184
Table 5-x. Correlations between Forms 2 and 3 of Officer Qualification Test and final average grade (Women's Reserve Midshipmen's School).....187
Table 6-x. Correlations between Forms 2 and 3 of Officer Qualification Test and components of final average grade (Women's Reserve Midshipmen's School).....188
Table 7-x. Means and standard deviations of grades in Basic Indoctrination, Communications, and Advanced Indoctrination (Women's Reserve Midshipmen's School).....188
Table 8-x. Average correlations between subtests of Officer Qualification Test and course grades (Women's Reserve Midshipmen's School)..189
Table 9-x. Simple vs. multiple correlation between Officer Qualification Test and course grades (Women's Reserve Midshipmen's School)..190
Table 10-x. Correlations between selection measures and final academic average (Reserve Midshipmen's School [Deck]).....193
Table 11-x. Scores on NROTC Selective Examination (Form C), for graduates and for academic failures (Reserve Midshipmen's School [Deck]).....196
Table 12-x. Scores on Officer Qualification Test, for graduates and for academic failures (Reserve Midshipmen's School [Deck]).....197
Table 13-x. Scores on separate sections of the Officer Classification Test, for graduates and for academic failures (Reserve Midshipmen's School [Deck]).....198
Table 14-x. Raw scores on Test N-4 (V-12 Comprehensive Objective Test), for graduates and for failures (Reserve Midshipmen's School [Deck]).....199
Table 15-x. Biserial correlation between test scores and the criterion of graduation vs. failure—together with product-moment correlations between test scores and final academic average (Reserve Midshipmen's School [Deck]).....200
Table 16-x. Detailed correlations between Test C-1 and Test N-4 (V-12 College Training Program).....204
Table 17-x. Comparison of multiple and simple correlations between individual parts of Test N-4 and sections of Test C-1 (V-12 College Training Program).....205
Table 18-x. Correlations between C-1 scores and average grade (V-12 College Training Program).....207
Table 19-x. Means and standard deviations of C-Test scores, N-Test scores, and average grades (V-12 College Training Program)....208
Table 20-x. Correlations between criteria (N-Test scores and average grades). (V-12 College Training Program).....209
Table 21-x. Average N-Test score and average grade for civilian-origin and fleet-origin cases (V-12 College Training Program).....210
Table 1-xi. Correlation coefficients between scores on aptitude tests and success in training in fourteen officer training programs.....224
Table 2-xi. Means and standard deviations of scores on aptitude tests for officers enrolled in fourteen officer training programs.....225

Table 1-XII. Group 1 Schools: Correlation coefficients (raw and corrected for restriction in range) between scores on six tests of the Basic Test Battery and final school grades for graduates of fourteen types of elementary Naval Training Schools....241
Table 2-XII. Group 1 Schools: Means and standard deviations on six tests of the Basic Test Battery for graduates of fourteen types of elementary Naval Training Schools....242
Table 3-XII. Group 1 Schools: Correlation coefficients between sums of scores on two tests of the Basic Test Battery and final school grades for graduates of fourteen types of elementary Naval Training Schools....244
Table 4-XII. Group 2 Schools: Correlation coefficients (raw and corrected for restriction in range) between scores on six tests of the Basic Test Battery and final school grades for graduates of five types of elementary Naval Training Schools....246
Table 5-XII. Group 2 Schools: Means and standard deviations on six tests of the Basic Test Battery for graduates of five types of elementary Naval Training Schools....246
Table 6-XII. Group 2 Schools: Correlation coefficients between sums of scores on two tests of the Basic Test Battery and final school grades for graduates of five types of elementary Naval Training Schools..247
Table 7-XII. Group 3 Schools: Correlation coefficients (raw and corrected for restriction in range) between scores on six tests of the Basic Test Battery and final school grades for graduates of two types of elementary Naval Training Schools....248
Table 8-XII. Group 3 Schools: Means and standard deviations on six tests of the Basic Test Battery for graduates of two types of elementary Naval Training Schools....248
Table 9-XII. Group 3 Schools: Correlation coefficients between sums of scores on two tests of the Basic Test Battery and final school grades for graduates of two types of elementary Naval Training Schools..249
Table 10-XII. Correlation coefficients (raw and corrected for restriction in range) between scores on Form 1 of the Special Tests of the Basic Test Battery and final grades for graduates of seven types of elementary Naval Training Schools....249
Table 11-XII. Means and standard deviations on Form 1 of the Special Tests of the Basic Test Battery for graduates in seven types of elementary Naval Training Schools....250
Table 12-XII. Comparative validity of single tests and combinations of tests of the Basic Test Battery in six types of elementary Naval Training Schools....250
Table 13-XII. Distribution of trainees in nine types of elementary Naval Training Schools according to quality classification and success in training....251
Table 14-XII. Distribution of trainees in eight types of elementary Naval Training Schools according to age....253
Table 15-XII. Correlation coefficients (product-moment and biserial) between age and final grade for trainees in eight types of elementary Naval Training Schools....254

Table 16-XII. Distribution of trainees in eight types of elementary Naval Training Schools according to amount of civilian education in years 255
Table 17-XII. Correlation coefficients (product-moment and biserial) between amount of civilian education in years and final grades for trainees in eight types of elementary Naval Training Schools... 256
Table 18-XII. Percentage of failure among trainees of ten types of elementary Naval Training Schools whose scores on the General Classification Test were (1) below the cutting score; and (2) above the cutting score, classified according to reasons for failure..... 257
Table 1-XIII. Means and standard deviations of available test scores and other selected personal data for trainee samples from twelve advanced Naval Training Schools for Enlisted Personnel..... 270, 271
Table 2-XIII. Correlation coefficients of school success with test scores and other selected personal data for samples from twelve advanced Enlisted Schools 274
Table 3-XIII. Criterion score means and standard deviations for four age groups in samples of trainees from ten advanced Enlisted Schools.. 276
Table 4-XIII. Relationship of prior Navy training to success in advanced service schools. (1) Biserial correlation coefficients between criterion and graduation or non-graduation from an elementary Naval Training School. (2) Product-moment correlation coefficients between criterion and grades in prior training..... 276
Table 5-XIII. Criterion score means and standard deviations of samples from ten advanced Enlisted Schools, subdivided according to civilian occupation groups..... 277
Table 1-XV. The relative contribution of each part-grade to the total variance of the composite final grade for graduates of two classes in a Basic Engineering School..... 305
Table 2-XV. Achievement of recruits at four Naval Training Centers as indicated by means and standard deviations on the Recruit Training Final Achievement Examination, Form 1..... 312
Table 1-XVI. Means and standard deviations for Pre-Radar Final Achievement Examination—first administration of experimental forms 322
Table 2-XVI. Performance on the CIC Final Achievement Examination of student officers receiving eight weeks of instruction and sea-experienced officers receiving four weeks of instruction at Tactical Radar School 324, 325
Table 3-XVI. Performance of students in two Reserve Midshipmen's Schools on Parts 1 and 2 of the Reserve Midshipmen's Schools' (Deck) Standardized Examination..... 326
Table 4-XVI. Intercorrelations among grades comprising the Final School Grade for the fifth class at the Tactical Radar School..... 327
Table 5-XVI. Intercorrelations among the grades comprising the Final School Grade for the ninth class at the Tactical Radar School... 328
Table 6-XVI. Correlation coefficients between scores on the Officer Classification Test and criterion measures at the Tactical Radar School, for classes on which the bases of grading differed..... 329

Table 1-XVII. Reliability coefficients, average item-test correlation coefficients, and average difficulty values for EE and RM Final Achievement Examination 337
Table 2-XVII. Part-whole and interpart correlation coefficients for EE and RM Final Achievement Examination, Form 6 338
Table 3-XVII. Correlation coefficients between scores on EE and RM Final Achievement Examination and marks at end of first month of Radio Materiel School 338
Table 1-XVIII. Distribution of different types of items used in the Advancement Examinations, Books I, II, and III 350
Table 2-XVIII. Codes for examination subjects used in coding of items in the McBee Keysort System as set up by the Test and Research Section 352
Table 1-xx. Distribution of sample population by ship type and by rating 382
Table 2-xx. Mean score on three selection tests, mean age, and mean number of years of civilian education by ship type and by rating for the sample population 395
Table 3-xx. Months of experience in rating by type of ship for six ratings: means and standard deviations of distribution 396
Table 4-xx. Average within-group correlation coefficients between months of experience in rating and unadjusted technical competence average scale values for six ratings 396
Table 5-xx. Average within-group correlation coefficients between Basic Test Battery scores and adjusted criterion values for six ratings 398
Table 6-xx. Comparison of the correlation coefficients between scores on three tests of the Basic Test Battery and the criterion for six ratings, computed by the combined and average within-group procedures 399
Table 7-xx. Intercorrelations, means, and standard deviations of selected variables computed for all cases in each rating for which complete data on these variables were available 400
Table 8-xx. Average within-group correlation coefficients between age in years and adjusted criterion values for six ratings 401
Table 9-xx. Average within-group correlation coefficients between years of civilian education and adjusted criterion values for six ratings 401
Table 10-xx. Means and standard deviations of scores on three Basic Test Battery tests for school, non-school, and combined groups for six ratings 402
Table 11-xx. Means and standard deviations of adjusted criterion measures for school, non-school, and combined groups for six ratings 402
Table 12-xx. Biserial correlation coefficients between school vs. non-school training and adjusted criterion values for six ratings 403
Table 13-xx. Significance of differences in mean adjusted criterion values for school and non-school trained groups, equated on mean scores of test of Basic Test Battery correlating highest with the criterion 404
Table 14-xx. Average within-group correlation coefficients between two measures of school success and adjusted criterion values for six ratings 404

FIGURES

Figure 1-I. Bureau of Naval Personnel: Organization as of July 15, 1945 5
Figure 1-VI. Sample Block Counting Item 58
Figure 2-VI. Sample Mechanical Comprehension Item 59
Figure 3-VI. Sample Surface Development Item 59
Figure 4-VI. Sample Tool Relationship Items 60
Figure 5-VI. Monthly trends in scores on Verbal and Mathematical Tests of Basic Test Battery 66
Figure 6-VI. Monthly trends in scores on Mechanical Tests of Basic Test Battery 68
Figure 7-VI. Literacy Test Sample Items 79
Figure 8-VI. Non-Verbal Classification Test Sample Items 81
Figure 1-VII. Profiles of mean scores of three groups on all sections of Officer Classification Test 105
Figure 1-IX. Enlisted Personal Inventory, Form 2, Part 1: percentage of normal and maladjusted men in Sample 1 tested at Receiving Station A who, at various score levels, would be referred for psychiatric examination 145
Figure 2-IX. Enlisted Personal Inventory, Form 2, Part 2: percentage of normal and maladjusted men in Sample 1 tested at Receiving Station A who, at various score levels, would be referred for psychiatric examination 147
Figure 3-IX. Enlisted Personal Inventory, Form 2, Part 1: percentage of men (well-adjusted, doubtful, and discharged), in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination 150
Figure 4-IX. Enlisted Personal Inventory, Form 2, Part 2: percentage of men (well-adjusted, doubtful, and discharged), in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination 151
Figure 5-IX. Enlisted Personal Inventory, Form 2, Part 1 plus Part 2: percentage of men (well-adjusted, doubtful, and discharged), in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination 152
Figure 6-IX. Billet Qualifications Blank, Form X-2(M), Scale N: percentage of men (well-adjusted, doubtful, and discharged), in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination 155
Figure 7-IX. Billet Qualifications Blank, Form X-2(M), Scales N plus L: percentage of men (well-adjusted, doubtful, and discharged), in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination 156
Figure 8-IX. Billet Qualifications Blank, Form X-2(M), Scale N: percentage of men (well-adjusted, doubtful, and poorly adjusted), tested at Receiving Station C who, at various score levels, would be referred for psychiatric examination 159

Figure 9-IX. Billet Qualifications Blank, Form X-2 (M), Scales N plus L: percentage of men (well-adjusted, doubtful, and poorly adjusted), tested at Receiving Station C who, at various score levels, would be referred for psychiatric examination....160
Figure 10-IX. Experience Comparison Index, Form X-1: percentage of men (well-adjusted, doubtful, poorly adjusted) tested at Receiving Station C and of neuropsychiatric cases, tested at a Naval Hospital who, at various score levels, would be referred for psychiatric examination....162
Figure 11-IX. Personal Check List, Form X-4 (27 Item Key): percentage of normal and maladjusted men in Sample 1 tested at Receiving Station A, who at various score levels, would be referred for psychiatric examination....165
Figure 12-IX. Personal Check List, Form X-4 (27 Item Key): percentage of normal and maladjusted men in Sample 2 tested at Receiving Station A who, at various score levels, would be referred for psychiatric examination....166
Figure 1-XV. Sample pencil and paper test items in Basic Engineering..296
Figure 2-XV. The 20mm. Performance Test being administered to trainees in Gunner's Mate School....296
Figure 3-XV. Card used in the administration of the 40mm. Identification Test to trainees in Gunner's Mate School....297
Figure 4-XV. Two gages used in rating products of shop work in Basic Engineering Schools....297
Figure 5-XV. Sample check-sheet used by proctor in administration of Direction Finder Operation Test (Model DAE)....302
Figure 6-XV. Improvement of scores in Torpedoman performance test (Assembling Turbine Bulkhead Leads) in successive classes at a Torpedoman School....304
Figure 7-XV. Improvement of scores in Small Arms Identification Test in successive classes at a Gunner's Mate School....304
Figure 8-XV. Prediction of success in Basic Engineering School by use of Basic Test Battery (before and after introduction of Achievement Testing Program)....307
Figure 9-XV. Prediction of success in Torpedoman School by use of Basic Test Battery (before and after introduction of Achievement Testing Program)....308
Figure 10-XV. Comparison of performance on Torpedoman Identification Test of two classes....308
Figure 11-XV. Comparison of classes in two Torpedoman Schools as to performance on Torpedoman Final Achievement Examination..309
Figure 12-XV. Comparison of performance of classes in two Basic Engineering Schools before and after training, measured by a standardized Final Achievement Examination....310
Figure 1-XVI. Sample pencil and paper items in the Pre-Radar Field..321
Figure 1-XVII. Group gains in scores on EE and RM Final Achievement Examinations, Form 1 Revised....339
Figure C, Appendix D-2. Sample Identification Test Card, with Part..481

PART I

THE NAVY'S SELECTION, CLASSIFICATION AND TRAINING PROGRAMS

CHAPTER I

PERSONNEL RESEARCH AND TEST DEVELOPMENT IN THE BUREAU OF NAVAL PERSONNEL: HISTORY AND SCOPE OF THE PROGRAM

WHEN war came to the United States in December 1941, the Navy faced the biggest personnel problem in its history. Because the Navy had remained relatively small during World War I, it had not been necessary at that time to establish an extensive personnel program. Procuring, classifying, training, and assigning to duty the four and one-half million officers and men required to man the ships and shore stations in World War II called for modern personnel techniques. In order to avail itself of the experience of government, education, and industry in personnel procedures, the Navy sought the assistance of various outside agencies and reorganized its internal organization in order to apply up-to-date techniques to its personnel problems. A direct result was the establishment within the Bureau of Naval Personnel of a unit of personnel trained in psychology which became the Test and Research Section. The development of this Section, its mission and organization, are the subject of this chapter.

MISSION OF THE TEST AND RESEARCH SECTION. The Test and Research Section (since January 1946 the Test and Research Division) was established to develop tests and carry on research studies designed to assist in selecting, classifying, and training officers and enlisted personnel from the time they are examined for admission to the Navy, through their indoctrination or basic training and specialized technical preparation, until they are satisfactorily assigned to duty aboard ships or at shore stations. Specifically, the Section's mission has been:

1. To construct, evaluate, recommend, and approve the adoption of psychological and educational tests and other instruments for use in determining the qualifications of applicants for admission to the naval service; in classifying personnel for technical training and for types of duty assgnments; and in measuring achievement in naval training programs;

2. To conduct research on the effectiveness of tests and other instruments used in selection, classification and training programs; on the effectiveness of selection and classification procedures; on the effectiveness of training; and on the performance of naval personnel in their duty assignments.

ADMINISTRATIVE ORGANIZATION OF THE TEST AND RESEARCH SECTION. The Bureau of Naval Personnel is responsible for the procurement, education, training, discipline, promotion, and distribution of officers and enlisted personnel of the Navy, except for the professional education of officers, nurses, and enlisted men of the Medical Department. From November 1942 through October 1945, the period during which the studies reported in this volume were undertaken and in large measure completed, the personnel test development and research program was developed in the Training Activity which has cognizance over naval training.

The administrative organization of the Test and Research Section in relation to the Bureau is shown in Figure 1; the organization and staff of the Section is shown in Appendix A-1. The responsibilities of the four Units in the Section have been as follows:

Selection Test Unit. Development and validation of aptitude tests for officers and enlisted personnel; and advancement in rating examinations for enlisted personnel.

Achievement Test Unit. Construction and evaluation of achievement examinations and other devices to measure the proficiency of officers and men in training programs.

Research Unit. Conduct of research on problems of classification and training of officer and enlisted personnel.

Radio Materiel Unit. Development of a coordinated program of test construction and research on selection, classification, and training of radio technicians.

SCOPE AND SERVICES TO NOVEMBER 1945. The work of the Test and Research Section has originated either as projects initiated by the Section or in response to requests made by Navy activities for assistance on specific problems. The extent of the use of tests and results of research studies and the technical services rendered to Navy programs may be summarized as follows:

To the Bureau of Naval Personnel and its Field Activities:

1. The Officer Procurement Division, the Officer Distribution Division, the Enlisted Recruiting and Induction Division, and the Enlisted Distribution Division have made use of selection and classification tests and the results of research studies in establishing and revising selection and classification procedures.

2. Reserve Midshipmen's Schools, Naval Reserve Officer Training Corps Units, Naval Training Centers and Naval Training Schools, and other training programs for officers and enlisted personnel have used achievement examinations for evaluating performance of individual trainees and efficiency of training.

3. Instructional and classification staffs have been given technical assistance in techniques of using test results, in techniques of developing

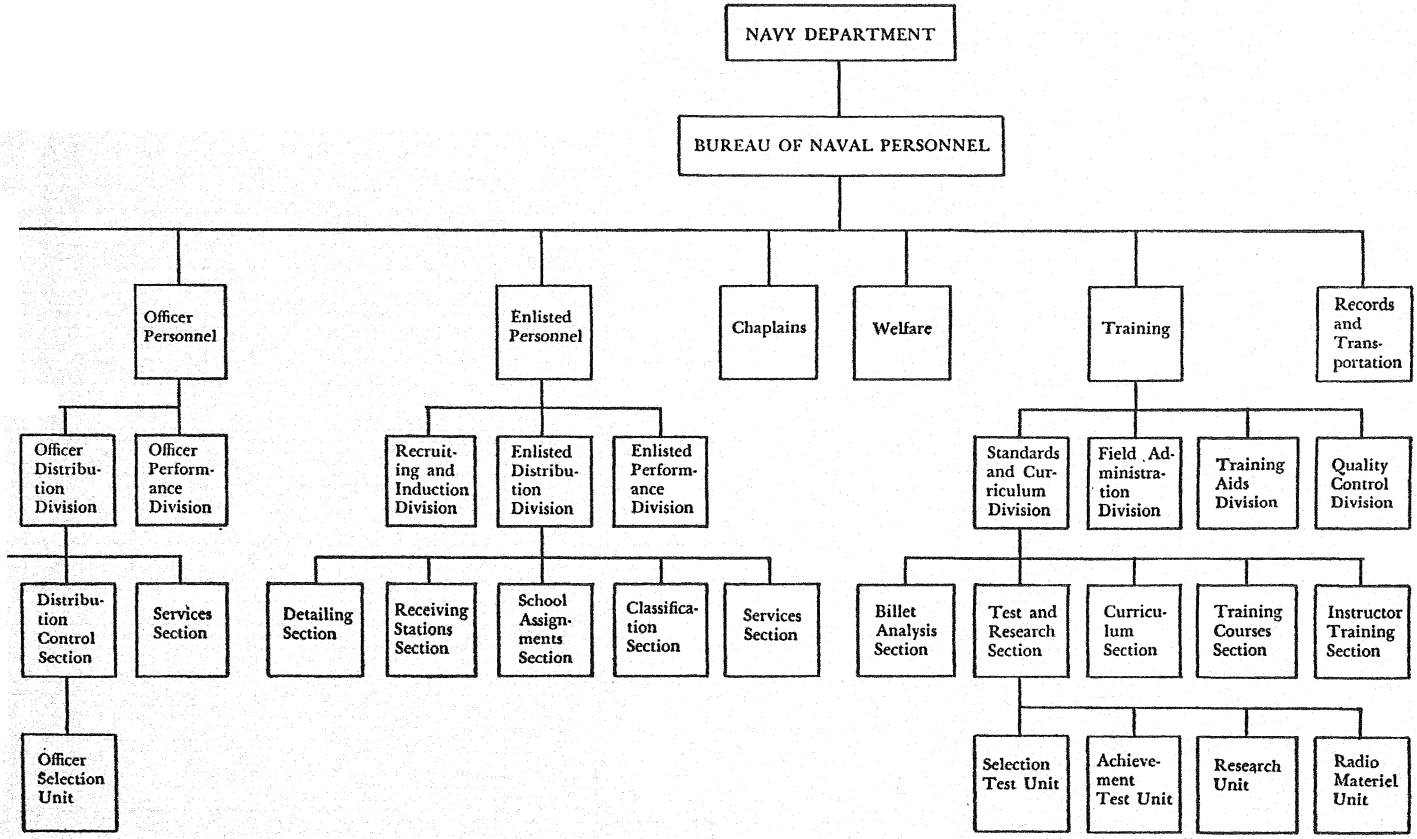


Figure 1-1. Bureau of Naval Personnel organization as of July 15, 1945.

tests for specific situations, and in development and improvement of grading systems.

4. The Educational Services Section has used the results of research studies in evaluating the program of off-duty education and orientation.

To Other Personnel and Training Programs:

1. The Training Commands of the Atlantic and Pacific Fleets have made use of the Section's tests and research findings in fleet schools, team training centers, and pre-commissioning centers.

2. The Amphibious Training Commands of the Atlantic and Pacific Fleets have used tests and the findings of research studies in their classification and training programs.

3. The Fleet Commands have used the fleet edition of the Basic Test Battery in classifying enlisted personnel for types of duty and assignment to advanced service schools.

4. At the request of the Office of the Commander-in-Chief, the Section has conducted research on the performance of personnel aboard ship, in order to improve selection, classification, and training.

5. Examining Boards on ships and at shore stations have used advancement in rating examinations in determining eligibility of enlisted personnel for petty officer ratings.

Personnel Testing and Research in the Navy before November 1942

Although psychological tests were in use in the Navy as early as 1912, there is no record of an organized program for testing naval personnel prior to 1924, when the Training Division was established in the Bureau of Navigation. At that time a General Classification Test was used in training stations to select enlisted men for Navy schools. Later this test was given at recruiting stations; as a result, men were eliminated at the recruiting level rather than after they had enlisted in the Navy. Other tests of special aptitudes were also introduced. Sometime before 1941 the use of these tests was restored to the recruit training activities. In December 1941 the following tests were in general use: General Classification Test, Mechanical Aptitude Test, Arithmetic Test, English Test, Spelling Test and Radio Aptitude Test.

Useful as these tests had been in peacetime when the Navy could select among applicants for enlistment, they were not adequate as the pressure of filling school quotas from limited sources became great. Throughout several months before and just after December 1941 many persons on ships and stations and at training installations became acutely aware that not all the men being sent from recruit training to naval training schools were capable of being trained for

their specialties. They were also discovering that the tests in use for selecting these men did not differentiate, for example, between good candidates for radioman training and good candidates for storekeepers schools. As a result, "home grown" testing programs developed at many establishments, and from a number of sources requests for and comments upon testing programs converged upon the Training Division in the Bureau of Navigation (after 13 May 1942 the Bureau of Naval Personnel), and upon the Office of the Commander-in-Chief of the United States Fleet. In April and May 1942, through correspondence and conferences, some of the specific problems regarding testing became focused, and request was made on May 2 by the Bureau of Navigation to the Office of Scientific Research and Development for advisory assistance in the solution of the Navy's testing problems.

For several months thereafter, two parallel approaches toward improving the Navy's testing program were developing, one within the Training Division in the Bureau of Naval Personnel and the other on a contract basis under the National Defense Research Committee.

THE PROGRAM IN THE TRAINING DIVISION TO NOVEMBER 1942. Throughout the summer and early fall of 1942 the Bureau of Naval Personnel was coming to grips with some of the immediate personnel problems facing the expanding Navy. Training, always a major activity of the Navy, loomed large as increasing numbers of untrained personnel were procured, and as the demand for skilled technicians in many branches multiplied rapidly. Training programs for enlisted men increased. It became more and more imperative to be able to assign recruits to training schools so that there would be the least possible attrition. As a result, a survey of the existing Navy tests, of Army tests, and of commercial tests which might be adequate for selecting enlisted men for naval training schools was instituted. The procurement of a large number of men from civilian life to be commissioned as officers presented a similar need for the development of a test to be used at Offices of Naval Officer Procurement to assess the qualifications of these men.

An increase in personnel in the Bureau of Naval Personnel was necessitated to carry on these activities, and during the summer of 1942 several psychologists were commissioned to assist the Civilian Educational Advisor, who until June 1942 had been the single professional expert on test construction in the Bureau. Work was thus begun on the development of an officer test, on the revision of the existing enlisted battery of tests, and on the study of tests for selection of personnel for special programs in sound schools and radio schools.

THE CONTRACT PROGRAM TO NOVEMBER 1942. Following the request in May of the Chief of the Bureau of Navigation for advisory assistance with the Navy's testing program from the Office of Scientific Research and Development, a series of conferences between naval personnel and members of various research and scientific groups was held. Army personnel were also contacted. In June 1942, the *Committee on Selection and Training of Service Personnel* was set up in the National Research Council, financed under a contract between the National Research Council and the National Defense Research Committee. As stated in the minutes for June 20, 1942, the Committee, organized as OSRD Project N-100 (later to become the Applied Psychology Panel), was to be "informal, quick, and direct, of as much use to the Navy and Army as possible, and would work always with the understanding that it would do only those things which it is asked to do by the Navy and Army."

The first activity of the Committee was to familiarize itself with the problems of Navy training through a series of trips to Navy establishments. Following these trips, a number of recommendations for test construction and research were made on July 25, 1942 to the Chief of the Bureau of Naval Personnel.

In the meantime, early in June 1942, the College Entrance Examination Board had been requested to prepare an examination for use in procuring officers for the contemplated Women's Reserve, and work was begun on this examination. Preparation of a test for use in procuring enlisted women was also started. By the first of September these projects were approved and placed by the National Defense Research Committee under the Committee on Selection and Training of Service Personnel as Project N-106 (Research and Development of the Navy's Aptitude Program).

ACCOMPLISHMENTS TO NOVEMBER 1942. These two programs of study of the Navy testing and training problems, one within the Navy and the other under contract, continued somewhat independently of each other until the late fall of 1942.

By that time the personnel in the Navy program had increased; a number of psychologists had been commissioned and assigned to naval training programs to assist with selection and classification procedures; some studies had been made within the Navy of the existing Navy tests, a few commercial tests had been recommended for use in special programs; and work had continued in the Navy on the development of a short form of an officer test to be used at offices of naval officer procurement.

Under the contract program, women officers had been given a series of tests; one test had been developed for use in the procurement of enlisted women; one form of an Officer Qualification Test

was developed; and studies had been done on the reliability and validity of the existing enlisted tests and on a Naval Reserve Officer Training Corps Selective Examination in use in colleges and universities.

Administrative Organization: November 1942 to November 1945

THE INITIAL ORGANIZATION. When the Bureau of Naval Personnel was established in May 1942 to supersede the Bureau of Navigation, an organization of management counselors was hired to make a comprehensive survey of the scope and functions of the Bureau. Their recommendations, designed to render more efficient the Navy's personnel program, were presented on August 20, 1942; and for a number of weeks thereafter, personnel in the Bureau were in process of studying, evaluating, and implementing the recommendations.

By November the organization within the Bureau had been greatly modified. Within the Training Division of the Bureau, the Standards and Curriculum Section was established early in November, and made responsible for developing specific answers to three basic questions stated in the survey report:

(1) What are the requirements men must meet in order to be selected for training in various subjects?
(2) What should they be when they have finished training, i.e. precisely what qualifications should they have gained from the training?
(3) In the light of (1) and (2), what must they be taught?

Three units were established, each to prepare the answer to one of the above questions: *Selection Requirements Unit, Training Standards Unit, and Curriculum Unit.*

Within the Selection Requirements Unit, designed to answer the first question "by devising the best possible methods for the proper selection and classification of enlisted personnel and officers in the Navy", five groups were soon set up. Two of these groups were made responsible for developing general plans for the selection process and for implementation of these plans (Officer Selection Group and Enlisted Selection Group) [1]. Three Groups were established to provide tests and research data (Test Construction Group, Research Group, and Quotas by Grade Group).

The Test Construction Group was made responsible for the development of tests, the Research Group for studies of classification and

[1] In the spring of 1944 these two programs were transferred to Officer Personnel Activity and Enlisted Personnel Activity respectively. See Figure 1-1.

training problems, and the Quotas by Grade Group for establishing, for different types of training, quotas for enlisted men with varying levels of ability. NDRC Project N-106 was continued as a service to the Navy's testing and research program, and a number of other National Defense Research Committee projects on special phases of classification and training were later established.

GROWTH OF THE PROGRAM. As the program of the Standards and Curriculum Section developed, it became necessary to add units and to reorganize certain aspects of the organization. On August 24, 1943 the Test and Research Unit was established within the Section, with the staffs of the Test Construction Group and the Research Group forming the nucleus of the new Unit. Its program was to construct and validate classification tests for enlisted personnel and officers and to conduct any research necessary to improve the efficiency of the classification and training program.

On September 23, 1943, at the suggestion of Quality Control Division, Training, the Test and Research Unit was officially charged with the responsibility for developing final achievement examinations for naval training schools. By January 1944 an Achievement Test Group was organized and work was begun on examinations for schools for Yeomen, Storekeepers, Electrician's Mates, Quartermasters, and other specialties. In May 1944, a program was initiated to construct examination materials for use by boards in determining qualifications of enlisted personnel for advancement in rating. In the summer of 1944, preparation of achievement examinations for officer training programs was inaugurated. In December 1944, work commenced on an achievement examination for recruit training.

Throughout the war the training of radio technicians was one of the most strategic and technical in the Navy. In January 1944, the program of test construction and research on training for this rating was organized as the Radio Materiel Group in the Test and Research Unit.

On January 1, 1945, the Test and Research Unit became the Test and Research Section, in what was by this time the Standards and Curriculum Division in the Training Activity. Late in 1944, a series of special studies had been instituted to evaluate naval training programs. During 1945 this undertaking was expanded to include such projects as the analysis of attitudes of men in the amphibious forces, the extent of use and influence of the Navy's Educational Services and War Orientation program, the use by Navy personnel of the facilities of the United States Armed Forces Institute.

During 1945, also, attention was focused on the validation in the fleet of classification and training procedures; the construction and validation (in cooperation with the Bureau of Medicine and Surgery)

of instruments to detect combat fatigue and related psychological disabilities; and the development and use of tests and research results in the Navy literacy training program, in transferring reserve officers to the regular Navy, in standardizing instruction in Naval Reserve Officer Training Corp Units, and in selecting candidates for the Naval Academy at Annapolis.

The program of the Test and Research Section in November 1945 was thus directed toward the development of research studies and psychological and educational tests for reconversion and for peacetime personnel programs.

The Continuing Program

TRANSITION. Throughout the war period the greater part of the test development and personnel research program of the Bureau of Naval Personnel was carried on by the Test and Research Section. Research projects, however, were developed by other groups within the Bureau. Enlisted Billet Analysis Section, Standards and Curriculum Division, was responsible for the continuing study of the nature of Navy billets for enlisted personnel. The Officer Billet Analysis Section, Officer Personnel Activity, performed comparable research for officer billets. Field Research Section, Planning and Control Activity, undertook to conduct opinion polls and planning surveys. As the programs of personnel research developed, and as plans were being formulated for continuing into the peacetime Navy the essential aspects of personnel research and test development, it became apparent that the several research groups could function more efficiently through closer integration of their programs. Accordingly, on November 5, 1945, the four sections were combined to form the Research Division in Planning and Control. On January 4, 1946, Research was raised to the Activity level and the Test and Research Section, together with the other sections, was raised to Division level.

THE PEACETIME PROGRAM. Plans for a permanent research organization in the Bureau of Naval Personnel were officially approved in January 1946. The Research Activity will undertake a coordinated program of personnel research and test development centered around the major personnel problems of the Navy. It is contemplated that the overall mission of the Activity will be (1) to conduct studies on personnel, policy, techniques, and procedures, and on the assignment, evaluation, promotion or advancement, and morale of officer and enlisted personnel; and (2) to develop such psychological and educational tests and other instruments as may be necessary for the selection, classification, training, and evaluation of performance of Navy personnel. Within this organization will be subsumed the wartime functions of the Test and Research Section.

CHAPTER II

SELECTION AND CLASSIFICATION OF OFFICER PERSONNEL

BETWEEN the years 1941 and 1945 the Navy was confronted with the task of procuring nearly 300,000 officers to meet the needs of the expanded fleet. These officers were required to fill hundreds of different billets ashore and afloat, some very similar to civilian occupations, others very different. It was important that these men and women be procured and assigned for training and duty so that their education, experience, and aptitudes might be used effectively. The procurement and assignment of these personnel were the respective responsibilities of Officer Procurement Division and of Officer Distribution Division, Officer Personnel Activity.

The Procurement Process

COMMISSIONING FROM CIVILIAN SOURCES. Applications of civilians for direct commissioning in the Naval Reserve were processed by offices of naval officer procurement. Each applicant, upon coming to an office of naval officer procurement to make application, was required to complete a standard application form which yielded fairly complete information regarding his background and training. Applicants found to be initially qualified for consideration for commissioning were administered physical examinations and psychological qualification tests, and were interviewed by two officers.

The Officer Qualification Test was used as a screening device for evaluating applicants with a great variety of backgrounds and for determining the individual's ability to be of service to the Navy. Two interviews were given to all apparently acceptable candidates in order to obtain the judgment of two different individuals as to the applicant's potentialities. The primary purpose of the interview was to determine (1) the general qualifications of each applicant, and (2) where each applicant might be of most service to the Navy, in terms of the available billets and the needs of the service. Each applicant found to be acceptable on the basis of the physical examination, test results, and interviews was investigated through references, friends, employers, city and state police, or the Federal Bureau of Investigation.

If the applicant was found qualified for final consideration for one or more types of duty open to procurement at that time, his

record was forwarded to the Bureau of Naval Personnel together with recommendations as to the types of duty for which he appeared to be best suited. In the Bureau, records of recommended applicants were routed to special reviewers, each reviewing applications in the light of the individuals' qualifications for particular types of duty. Applications were rejected only when they had been examined by all reviewers who could possibly recommend appointment of the applicants. Applications found acceptable were routed, recommending appointment, to a final review board whose function was to make the final recommendation as to appointment or rejection.

COMMISSIONING OF ENLISTED PERSONNEL. The same opportunities afforded to civilians for commissions in the Naval Reserve were open to enlisted personnel. Enlisted personnel in recruit training were screened by classification officers and those found basically qualified for commissions had the opportunity to apply. Furthermore, any enlisted man at the end of a six-month period was given the opportunity to apply for a permanent commission in the Naval Reserve. Information similar to that required of civilians was required of enlisted applicants. Applications of enlisted personnel were reviewed in the Bureau of Naval Personnel in the same manner as were those from civilian sources.

The V-7 program also was open to enlisted personnel under thirty years of age who had completed two years of college. Personnel selected for V-7 training by the Bureau of Naval Personnel were ordered to midshipman training either directly or via a brief period of academic refresher training. In addition, commanding officers were requested to submit recommendations for temporary appointment of qualified enlisted personnel to commissioned rank. These recommendations were reviewed and approved or disapproved by an administrative board established in the Bureau.

V-12 PROCUREMENT.* V-12 personnel were selected from civilian sources to enter upon active duty on July 1, 1943, November 1, 1943, and July 1, 1944. Initial selection was made by means of nationwide tests administered in high schools and colleges throughout the country. Final selection of physically qualified applicants was made by state selection committees composed of business men, educators, and naval officers. Selection was based upon test scores, high school and college records, and the results of interviews given at offices of naval officer procurement. Enlisted personnel were also selected for V-12 training by commanding officers of ships and shore stations in accordance with procedures outlined in Bureau directives. Insofar as possible, the same selection requirements outlined for civilians were followed in selecting V-12 candidates from enlisted ranks.

V-1 AND V-7 PROCUREMENT. Prior to the entry of this country into the war, the V-7 program was established. The V-7 program at that time was open to candidates who had successfully completed two years of college. Following induction, officer candidates were subject to call to active duty for training cruises and were eligible for commissioning upon completion of college. Following the beginning of the war, the training cruise phase of the program was eliminated and V-7 trainees were called, upon graduation, to active duty for midshipman training in reserve midshipmen's schools, at the end of which they were commissioned.

The V-1 program was opened in early 1942 to freshmen and sophomores as a companion program to the V-7 program. Personnel enlisted in the V-1 program were to be retained on inactive duty until graduation, at which time they were called to active duty in midshipman training status. The larger part of these enlistees were transferred to V-12 upon the opening of that program.

Selection for both the V-1 and the V-7 programs was based upon physical fitness, academic record, and "officer-like" qualities. Students in these programs were required to complete certain required courses prior to graduation from college and subsequent admission to reserve midshipmen's schools.

Classification of Officers

When the officers had been procured, it was the task of the Officer Distribution Division, Officer Personnel Activity, of the Bureau of Naval Personnel so to distribute them as to make maximum use of the skills and experience which they already possessed. Officers were urgently needed to man the ships which were sliding down the ways in a production program of unparalleled scope. Time for training them was short; consequently each officer had to be placed so that, insofar as possible, his civilian education and experience could take the place of the extensive naval training which he would have received in peacetime to prepare him for his assignment. In the paragraphs that follow, a brief description is presented of the selection and classification process.

CLASSIFICATION AT RESERVE MIDSHIPMEN'S AND INDOCTRINATION SCHOOLS. The Officer Selection Unit (first established early in 1943 in the Standards and Curriculum Division, Training Activity, and transferred in early 1944 to the Officer Distribution Division, Officer Personnel Activity) was immediately concerned with the graduates of Navy indoctrination and reserve midshipmen's schools. These new officers, about to embark upon their first Navy assignments, were ideal subjects for classification. Interviewing officers were as-

signed to each indoctrination and reserve midshipmen's school to assist the Bureau of Naval Personnel in placing the graduates. The interviewers themselves were carefully chosen. Many of them had graduate degrees in psychology, and all of them had had extensive pre-Navy experience in personnel work in industry or in colleges and universities.

These interviewers were first given a training course under the direction of the Officer Selection Unit. They studied the system of officer detailing by which the Bureau of Naval Personnel distributes all naval officers. They were coached in the program of testing, interviewing, and recommending which was to be standard in every midshipmen's and indoctrination school. And finally they went on a field trip to various naval bases so that they might learn first hand about duty aboard various types of ships and stations. During this trip they "hit the beach" in a practice landing from an LSM, made a run in a PT boat, and spent a day on an anti-submarine patrol. They went aboard dozens of ships of many types and visited the classrooms and laboratories of advanced training schools of various kinds. At the end of this training, the interviewing officers took up their posts at the indoctrination and midshipmen's schools, prepared for the task of selecting new officers for their first naval duties.

The selection program at these schools fell roughly into three sections: (1) acquainting the new officers with the kinds of naval duty open to them, (2) evaluating each officer's qualifications and preferences for duty through testing and interviewing, and (3) making recommendations to the Detail Section, Officer Distribution Division, Officer Personnel Activity, Bureau of Naval Personnel, for the distribution of each graduating class.

The interviewing officer introduced the subject of naval duty to the midshipmen or indoctrinees in a "billet lecture". Here he briefly described the assignments afloat or ashore in which the new officers might be placed, paying particular attention to those spots where officers were most urgently needed at the moment. The pattern of officers in demand in certain duties changed constantly with the progress of the war, so that the billet lecture for one indoctrination class might differ considerably from that for the next month's class. The student officers had to be prepared for the possibility that even in the time between the billet lecture and the assigning of the class, the exigencies of the war might eliminate certain needs and create others. The interviewing officer tried always to foster in the students the realization that their individual aptitudes and qualifications had to be utilized in relation to prosecuting the war. A man might have had years of successful experience as a trial lawyer, yet at a time when the amphibious program was desperately in need of officers,

his one or two summers of small boat handling might have far more significance.

The second and perhaps most important function performed by the interviewing officer was making an evaluation of each officer's best potential contribution to the Navy. This was accomplished by testing and by interviewing. Every officer was given the Officer Classification Test, a four-part battery designed to measure verbal facility, mechanical aptitude, mathematical competence, and ability to visualize spatial relations. Those officers who appeared to be candidates for such specialized training as radar materiel or tactical radar were given the selection tests developed for these types of training. The construction and validation of these tests is described in Chapter VIII.

Testing took place before interviewing, so that at the time of the interview each candidate's test scores were before the interviewing officer. These scores, together with the officer's background, civilian or naval, his indoctrination or reserve midshipmen's school marks, and his own choice among the assignments open to the class, together with his appearance, physique, and personality made up the general picture from which the interviewer made his estimate of the person's abilities.

The interviewing officer's recommendations to the Bureau were the result of matching the individual sets of aptitudes to the jobs open to the class. To aid him in this task the interviewing officer had at hand the *Billet Selection Requirements Manual,* compiled by the Officer Selection Unit. This manual which was continually in process of revision listed some one hundred sea and shore assignments with the qualifications considered desirable or essential for each—educational background, civilian experience, minimum test scores, and personal qualities.

The interviewer's recommendations for assignment were transmitted to the Bureau on Duty Recommendation Forms. These forms contain a record of each individual's qualifications, grades, and test scores, a statement of his choice of duty, and the interviewer's recommendation. The information on these sheets was especially valuable in those instances where shifting war needs necessitated a distribution of officers different from that on which the interviewer had based his recommendations. The Duty Recommendation Form became a part of each officer's record, to be used at any time when he became available for transfer or when he requested change of duty.

The effect of this program at reserve midshipmen's and indoctrination schools was to reduce the time necessary for training officers for competent performance in their Navy billets. Likewise it kept to a minimum the number of officers who "bilged" from special

training courses. The contribution to officer morale made by considering each man's duty assignment individually, while it cannot be measured in tangible form, appears to have been considerable.

CLASSIFICATION AT OPERATIONAL TRAINING CENTERS. Interviewing and selection programs similar to those at indoctrination and reserve midshipmen's schools were carried out in many operational training activities such as the Destroyer Training School, the amphibious training bases, and the APA (attack transport) Pre-Commissioning School. In these schools, selection was on a different level; officers were interviewed and their grades and test scores evaluated for the purpose of placing each in the most appropriate post aboard ship. While testing was done at these centers, the primary emphasis was placed on evaluating the data previously gathered on each officer in relation to a specific billet assignment.

CLASSIFICATION AT NAVAL RESERVE OFFICER TRAINING CORPS SCHOOLS. The Navy college training programs such as V-12 and Naval Reserve Officer Training Corps trained a great many officers for commissioning. These units were located at colleges and universities all over the country and the separate groups were too small to warrant full-time interviewing officers for each. Interviewing and selection of these men for their first duties as naval officers were carried out by traveling interviewing officers from the Bureau of Naval Personnel. The students were given billet lectures, received informational pamphlets about naval duties, took the Officer Classification Test and various special aptitude tests, and were interviewed by the representative of the Bureau. The recommendations of the interviewer were sent to the Bureau on Duty Recommendation Forms, in the same manner as were recommendations for persons in midshipmen's and indoctrination classes.

CLASSIFICATION ABOARD SHIP. The classification of officers aboard ship was a relatively late development due to the fact that primary emphasis was placed on the processing of personnel to be assigned to training programs or new ships. In the spring of 1945, however, several requests for assistance in the classification of their officer personnel were received from commanding officers of ships being shifted from the European to the Pacific theatre of operations. Principally, the commanding officers of these ships were officers who had had tours of duty at naval training stations or other activities where the classification of officer and enlisted personnel was accepted as a standard procedure. In the shipboard programs the classification of officers was usually carried out at the same time as for the enlisted personnel. By V-J Day the officer personnel of more than twenty large ships had been processed. Included in a complete shipboard classification program were the following: administration of classifi-

cation tests; interviewing each officer; bringing up to date the Officer Qualifications Record Jackets; explaining the findings of the classification program to the executive officer or commanding officer or both; and making recommendations concerning the use of the data. Ideally, classification data should be obtained on all officers before they are assigned to duty, but since this had not been done, shipboard classification offered at least one opportunity to include officers in the program who had not been processed before assignment to duty.

INTERVIEWING OF OFFICERS RETURNING FROM SEA FOR REASSIGNMENT. In the spring of 1945, classification-interviewing was instituted at the ports of entry through which officers passed on returning from sea duty. Each officer was given a comprehensive interview with an experienced interviewing officer, during which his qualifications for next duty were reviewed. During the interview he could ask questions about the availability of billets in which he was interested and obtain first-hand information on Navy practices with respect to rotation and leave. In turn, the interviewing officer asked him questions necessary to obtain a complete summary of his naval and civilian skills and interests.

A Duty Recommendation Form was used for transmitting the interviewer's recommendations to the Bureau. This Duty Recommendation Form was a variant of the one used at naval training activities, providing more space for listing naval experience, and eliminating the space for recording test scores and grades.

In contrast to the fairly precise quotas for distribution of the officers which were available for interviewers at reserve midshipmen's and indoctrination schools, interviewing officers at ports of entry had no specific quotas to guide them. They were, however, supplied with monthly reports of sea billets open, which indicated the acuteness of the needs in certain areas, and with monthly reports of the shore jobs open, with the qualifications required in each.

Having familiarized himself with these reports from the Bureau, the interviewer was able to assist the returning officer to state his duty preference in the light of the billets available. The interviewers were, in fact, in such close touch with the Bureau of Naval Personnel, that the time spent by returnees in the port of entry office was the equivalent for each of a personal visit to the Bureau. The interviewer could prepare the returnees in general for what they could expect in their next assignments; could explain the ramifications of rotation practices as they applied to different types of duty; and could effectively stop many of the fallacious rumors with which returning officers were frequently familiar. This individual handling of returning officers was designed to improve their attitude toward

their next assignments and to insure their better performance in whatever jobs they were placed.

Officer Classification Records

Because of the speed with which the Navy had to be expanded, there was no time for selection and classification activities to be established in the early months of the war. Hence many naval officers were assigned to duty where the need was greatest, with little regard for their preferences or qualifications. This was inevitable. But these officers too, and in fact all naval officers, were eventually covered by the International Business Machines punched card classification and records system.

The basis of the I.B.M. punched card system is the Officer Qualifications Questionnaire which each officer fills out and files with the Bureau. This questionnaire lists all his important qualifications—foreign language facility, foreign travel, education, civilian experience, and naval duties. The qualifications of each officer are transposed into numerical code which is punched into cards. The cards then become an index to the questionnaires and can be drawn upon whenever an officer with a specific set of qualifications is required for a special job. The card system also serves as a medium for making officer personnel surveys and studies upon which to base plans and policies.

One copy of the Officer Qualifications Questionnaire becomes the foundation of the Officer Qualifications Record Jacket. This jacket is carried by each officer from one duty station to another. Besides the questionnaire, which sums up his education, civilian work experience, and naval duties, the jacket contains a tear-off section from each fitness report submitted on the officer. This contains his name, file number, rank, classification, date of reporting, and a description of the duties performed since the last report. By means of this addition to the jacket, the record of the officer's qualifications is kept up to date. Another copy of the tear-off section of the fitness report is sent to the Bureau of Naval Personnel, where it is used to bring the punched card record up to date.

Impact of Officer Classification on Navy Personnel Procedures

Despite the late start of the officer classification program, it has had a significant effect upon the processing of officer personnel. As a result of the success of wartime officer classification, post-war plans for officer selection and classification include the following:

1. Continuation of an Officer Qualifications Unit in the Officer Distribution Division to carry on the various phases of an officer classification program.

2. Maintenance by the Officer Qualifications Unit of accurate records on all officers to the end that each officer will receive shipboard experience in accordance with the Navy rotational training program.

3. Assignment of a technically qualified district personnel classification officer to each naval district to carry on officer classification activity after the reserve officer personnel have been demobilized.

4. Sending of interviewing officers to the Naval Reserve Officer Training Corps Units and other training activities as needed to assist in the classification of officer personnel and officer candidates.

5. Maintenance of up-to-date qualification records on reserve officer personnel by means of questionnaires sent out every two years to all officers on inactive duty.

6. Administration, as needed, of officer selection and classification tests at training activities and other naval commands.

7. Continuation of the Officer Qualifications Record Jacket as the basic record for officer personnel.

By the establishment of the above administrative procedures and classification techniques in peacetime, the Navy will, in the event of an emergency, be able to mobilize its officer personnel in a minimum of time and assign them to billets where their abilities and experience can best be utilized.

CHAPTER III

SELECTION AND CLASSIFICATION OF ENLISTED PERSONNEL

In addition to the procurement and classification of the 300,000 officers described in Chapter II, the Navy had to recruit, train and assign approximately four million enlisted personnel. Furthermore, the rapid technological developments of modern naval warfare in such fields as gunnery, fire control, aviation, and electronics required that enlisted men be trained for and assigned to more than eight hundred different types of jobs varying in complexity from the deck hand who swabs, chips paint, and polishes bright work, to the radio technician who maintains and repairs highly technical and complex radar and electronics gear. The tasks of classifying men for the type of work in which they would be most likely to succeed, and of aiding personnel officers in assigning each man to the right job was the responsibility of the Enlisted Classification Section of the Enlisted Personnel Activity of the Bureau of Naval Personnel. At the time of V-J Day more than 1,100 carefully selected classification officers and classification interviewers (enlisted personnel) had been trained and assigned to provide classification services at more than one hundred different naval activities ranging in function from recruit training centers to forward area personnel distribution points.

Procurement of Enlisted Personnel

Enlisted personnel were procured during the war primarily through Selective Service and secondarily through the voluntary recruiting program administered through recruiting stations located in all large cities. Before the war, recruiting stations had been responsible for the procurement of all enlisted personnel. Applicants were required to meet both physical and intellectual qualifications. A General Classification Test was used in addition to an educational requirement in estimating the intellectual capacity of applicants.

During the war recruiting stations administered several special procurement programs in addition to carrying on their normal recruiting functions. In the radio technician recruiting program a special test, the Radio Technician Selection Test, was administered to all applicants. For most other special recruiting programs, such as for combat aircrewman, ship repair units and construction battalions, selection at the recruiting station level was accomplished by an examination of the applicants' educational and vocational ex-

perience. In the post-war recruiting program a psychological screening device, the Applicant Qualification Test, will be used in addition to the usual physical and psychiatric screening examinations.

Classification of Enlisted Personnel

CLASSIFICATION AT RECRUIT TRAINING COMMANDS. All personnel procured through Selective Service or recruiting stations were sent to recruit training commands (prior to 1944 naval training stations) for recruit training and for classification. Each classification office was under the direction of a classification officer, and all recruits were given a battery of classification tests and were interviewed by classification interviewers. The basic functions of classification in recruit training were: (1) to determine and record on standard forms, the aptitudes, skills and abilities that would indicate the type of naval duty for which each recruit was best fitted; (2) to recommend each recruit for the type of training or duty for which he was best qualified; and (3) to effect the assignment of each recruit to that type of training or duty for which he was best fitted by matching the man's qualifications with the Navy's needs as reflected in quotas issued by the Bureau of Naval Personnel.

These functions were carried out in accordance with the best available personnel techniques and methods. All recruits were first informed about available training schools and the duties of naval ratings by means of carefully planned lectures, films, and pamphlets. They were then given the Basic Test Battery for enlisted personnel.

These tests included the following: General Classification Test, Reading Test, Arithmetical Reasoning Test, Mechanical Aptitude Test, Mechanical Knowledge Test (Mechanical), Mechanical Knowledge Test (Electrical), Clerical Aptitude Test, Spelling Test, Radio Code Test.

The tests were administered under conditions as ideal as was possible in view of space limitations, and were scored, for the most part, by machine. Test scores were then checked and recorded by machine on each man's Enlisted Personnel Qualifications Card so that when the man appeared for interview, the interviewer had before him a complete picture of the test results. Aptitude testing was an essential feature of recruit classification, providing the most objective basis available for the classification and assignment of the man, particularly since a considerable proportion of inductees were fairly young, so that little opportunity had existed to acquire vocational skills or experience. Since test scores were recorded on the Qualifications Card, they could be used at any subsequent point in the man's naval career as a basis for evaluation and classification.

When recruits possessed previous civilian experience and training closely related to types of work available for them in the naval service, test scores were of secondary importance except as they indicated that the man was of such a low mental caliber that he could not be expected to learn quickly enough to qualify for the duties of a rating. In order to facilitate the evaluation and recording of the recruit's civilian experience, training, hobbies, and interests, a standard aid-to-the-interview blank was completed by each man prior to the interview. This form gave the recruit an opportunity to list his qualifications and to express his interests in various types of Navy jobs.

In the great majority of cases, it was a combination of a man's test scores, civilian work experience, motivation, previous training, and interests which guided the interviewer to a decision as to what types of duty the recruit was best qualified to perform. Duty recommendations were usually recorded in rather broad terms so that a man would not be too narrowly classified. This was necessary and desirable not only in the man's best interests but also because of variations in quotas from the Bureau of Naval Personnel which required flexibility in detailing.

To facilitate school assignments, which constituted a major phase of recruit classification, certain basic data from the Enlisted Personnel Qualifications Card were punched into I.B.M. cards for mechanical sorting so that selection could be rapid and accurate. In addition to the man's name, rating and service number, the I.B.M. selection card provided space for punching a civilian occupational code, an evaluation code and a first and second duty choice or recommendation.

Roughly speaking, about 40 percent of all recruits were selected for elementary naval training schools; about 10 percent were selected for special billets, immediate rating, or commissioning; and the remaining 50 percent were sent directly to ships or stations for duty as "general detail" hands. An adequate system for classifying this latter group was never fully developed and as a result many men who had useful skills were often not used effectively in their subsequent assignments. If an adequate classification system had been established earlier, it should have been possible to process the "general detail" population as well as the other groups.

CLASSIFICATION AT PRE-COMMISSIONING CENTERS. Many recruits and some elementary school graduates were sent directly to pre-commissioning training centers where they were assigned to new construction vessels. To assure continuity in the classification and assignment process it was necessary to establish classification centers at all such activities. The basic functions of classification at pre-

commissioning activities were to: (1) process and provide balanced crews for new construction vessels; (2) provide sufficient information about each man assigned to a ship to assure his proper utilization; and (3) assist the ship's executive officer or personnel officer with technical aids and methods that would assure an effective system of personnel administration aboard ship.

The majority of men detailed to new construction were not assigned to specific ships and this permitted the orderly scheduling and processing of all such men in order to assure each ship of its fair share of high, medium and low caliber men. This process known as "balancing the crew" was important, since otherwise one ship might have been assigned all the better qualified men and another might have received nothing but relatively poor ones. Aptitude test scores, previous civilian experience and training, physical qualifications, and the man's motivations were important factors in the balancing process. Accordingly, much use was made of the Qualifications Card which, as stated earlier, was usually prepared for recruits shortly after their induction. If the card was not available or was incomplete, necessary steps were taken to complete it before assignment.

In addition to this information about the man, it was also necessary to know the jobs available for the men aboard ship. Such information was secured by a review of the ship's complement and discussion with department and division heads, if they were available, as to their views on desirable qualifications for types of billets. The resulting job data were set up in card files so that the qualifications of the man and the job could be matched.

In order to man properly a ship's Watch, Quarters, and Station Bill, it was necessary to assign many men to jobs in gun crews and other battle stations on the basis of potential qualifications. To aid in this process, special selection devices such as the selectometer (developed by the National Defense Research Committee) and visual sorting devices making use of selective placement factors were developed and used with varying degrees of success. Special tests were also used in this process, including the Ortho-Rater tests for visual acuity and other visual capacities, the Telephone Talker Test for selecting men who would be most suitable as talkers on sound-powered phone circuits, and the Sonar Pitch Memory Test for identifying men who possessed sufficient pitch discrimination to qualify as sound gear operators.

The remaining functions of classification units at pre-commissioning centers were concerned with developing a useful system of personnel administration for each ship. This involved setting up an open Qualifications Card file, either centrally or by division depending upon the size of the ship. Special cross-index files also were

set up containing special data such as names of men possessing high test scores, college degrees, and special skills in foreign languages or experience as musicians or entertainers. These cross-index files were useful in filling special billets once the ship was commissioned. In addition, a cross-index file of men best qualified to "strike" (serve as apprentices) for petty officer ratings was established so that when vacancies occurred in any division, the most likely non-rated man on board could be selected and assigned for training to fill the vacancy. Finally, an officer and yeoman were trained in the use and interpretation of information contained in these files and in the technique of keeping them up-to-date.

The principal defects in classification work at pre-commissioning centers were the unavailability, and the lack of knowledge of qualifications of key members of the crew, who were usually sent to the yard where the ship was being built rather than to the pre-commissioning center; lack of specific complement and billet data required in order to assign men properly; and inadequacy of time permitted for classification and assignment of the crew which precluded the "pooling" of men with special abilities and qualifications.

Classification at Receiving Stations. At one time or another nearly every enlisted man in the Navy is ordered, in transient status, to a receiving station for assignment to a new ship or station. In order to assist in the evaluation of such men's qualifications, classification centers were established at virtually all continental receiving stations and at the Receiving Station, Pearl Harbor. The majority of men processed at receiving stations were assigned their next duty by the personnel office of the Commander, Service Forces, Atlantic Fleet; Commander, Service Forces, Pacific Fleet; and Commander, Western Sea Frontier. Accordingly, all men processed by classification centers whose ultimate assignment was controlled by the personnel office of the Fleet Service Forces, were reported on availability lists or cards to these offices. These lists or cards contained not only the man's name, rating, and pay grade, but also a brief summary taken from the Qualifications Card concerning the man's civilian and Navy experience and training, and a duty recommendation. In determining duty recommendations, the classification interviewers used the techniques and methods previously described, including, when necessary, testing, interviewing, and completion of records.

Receiving stations processed not only men going to sea for the first time but also sea-experienced men returning to the United States for leave and reassignment. The latter group constituted the largest single category of men processed by receiving stations and required a slightly different type of emphasis from that at other

activities. Billets were analyzed and interviewing aids developed to give interviewers a better understanding of the various jobs performed aboard ship. As a culmination of this emphasis, a *Manual of Enlisted Navy Job Classifications* was developed and published in October of 1945. Although too late to be of value during the war, this manual in preliminary form was used to good advantage in reporting men's qualifications and in reflecting the replacement requirements of ships.

Another function of classification centers at receiving stations was to contact ships in the general area served by the center in order to offer classification services similar to those at pre-commissioning training centers. Qualifications Cards were completed on all the members of the crew, and files were organized. Cross-indexes were developed and personnel were trained to maintain the entire system. Operating ships were also assisted in stating their personnel requirements in the form of replacement schedules. These were sent to the personnel offices of the Fleet Service Forces for use in detailing personnel.

The latter function was comparatively untried, since it involved use of the preliminary job classification titles and codes, but it proved effective in improving detailing and in achieving proper use of classification recommendations made on transient personnel.

Perhaps the most successful phase of classification at receiving stations was that concerned with the processing of men for advanced training schools. These men, subject to the assignment control of the Bureau of Naval Personnel, were selected by commanding officers of ships and stations for advanced training in accordance with quotas allocated to all fleet activities. Classification centers screened these men in order to determine whether or not they required the training to which they had been ordered, or, on the other hand, whether or not they would benefit from such training. It is estimated that substantial amounts of time and money were saved by this screening, because it was found that many men had been ordered to training who were already fully qualified to handle their duties on the latest types of equipment, while others were found who were of such limited ability that advanced training would have been of little value. Another advantage of such screening was that it aided the Bureau in determining the level of training to which each man should be ordered. For example, a motor machinist who had worked only on gasoline engine repairs would require both basic and advanced diesel school training, whereas a motor machinist who had worked on landing craft or small craft diesels would benefit most from advanced diesel training.

The development of an experience code and classification system

for this school screening program was the first step in the development of the *Manual of Navy Job Classifications*. This program, more than any other, emphasized the wide differences in experience and ability among men in the same rating and pay grade, and pointed out the necessity for a system of job classification more detailed than was provided by the existing rating structure.

Classification at Naval Districts and Shore Establishments. To assure maximum effective utilization of enlisted personnel in the Navy's wide-spread network of shore establishments, classification officers and enlisted personnel were assigned to all naval districts. The officers so assigned were called district personnel classification officers and were responsible for both officer and enlisted classification functions. In processing enlisted personnel, these officers made certain that every man coming into a district for assignment to duty was screened either by one of the interviewers assigned to him or by the nearest classification center.

Extensive work was also done to complete Qualifications Cards and cross-index files on all shore establishment personnel assigned to district activities. This was accomplished by sending squads of classification interviewers to process all members of ship's company at each district activity. The personnel officer and several assistants were then trained in the use of classification data for assignment and reassignment purposes. In order to implement this program at its inception, two large-sized units of interviewers were assigned to move from district to district and activity to activity in what was known as "The Flying Circus" or "Beach Survey". After this initial coverage of all districts from Boston to Pearl Harbor and from Chicago to Panama, district classification officers took over and maintained an active classification program with small units of interviewers assigned permanently to their staffs.

District personnel classification officers also acted as the field representatives of the Enlisted Classification Section of the Bureau of Naval Personnel. As such, they coordinated classification functions at all activities in the district including training centers, precommissioning centers, service schools, and receiving stations.

Classification at Service Schools. Studies at elementary service schools showed that there was an appreciable amount of attrition among recruits assigned to service school training. To remedy this condition, classification interviewers were assigned to counsel and advise service school trainees and to assign all graduates to the types of duty which would best utilize their skills and abilities. It was discovered that much of the attrition was due to disappointment with, or lack of interest in, school work on the part of the trainees. It was recommended to recruit training commands that in making

recommendations greater emphasis be placed on the interest factor. This was an improvement in selection procedures. Many students who otherwise might have dropped out, were counseled and assigned to other schools; others were stimulated and encouraged so that they found new interest in the school to which they had originally been assigned. Classification officers and classification interviewers at service schools, like those at other activities, also screened men for submarine, motor torpedo boat, and other types of special duty, by administering a personal inventory and by determining each man's general adaptability for such special programs.

CLASSIFICATION ABOARD SHIP. Officers and men especially trained to perform classification functions were assigned to only a few of the larger combatant ships. However, between 1,500 and 2,000 combatant and auxiliary vessels, ranging from patrol craft to battleships, were contacted and processed by classification centers. The nature of this service has already been outlined in the paragraphs describing classification procedures at pre-commissioning centers and receiving stations. Although it is difficult to evaluate the effectiveness of classification services provided to ships, it can be said on the basis of letters received from and discussion with executive officers and commanding officers, that many ships made practical use of classification services in setting up systematic personnel procedures of their own. The important point is that many officers who were and will be charged with the command of ships and stations have been convinced through their own experience of the value of a systematic personnel procedure for (1) analyzing the job to be done, (2) determining the qualifications of the men available, and (3) selecting on a scientific basis the man best qualified for the job.

Records for Enlisted Personnel

The basic record for all enlisted personnel is the Service Record. This contains, in addition to personal history data, the official record of the individual from the time he enters the service until he is separated. One copy accompanies the individual wherever he goes; a duplicate copy is kept in the Bureau of Naval Personnel. The record is kept up-to-date by the addition of new pages whenever there is a significant change in the individual's status, e.g. change in duty station, advancement in rating, or change in beneficiary. The first entries in the Service Record are made by the recruiting stations for voluntary recruits and by the induction centers for personnel procured through Selective Service. Official documents and correspondence are kept in a folder inside the cover page of the record.

In addition to the Service Record, there was prepared for recruits during World War II an Enlisted Personnel Qualifications Card on which were recorded significant personal data, test scores, record of vocational experience, schools attended, billets occupied, statement of hobbies and interests, and the classification interviewer's recommendation for duty assignment. This card could be placed in an open file for ready reference or left in the Service Record. The card is intended to give a brief concise picture of a man's qualifications. As stated earlier, the card is filled out by a classification interviewer, usually at a naval training center or advanced classification center.

The Impact of Enlisted Classification on the Navy Personnel System

The growing awareness of the value of systematic classification and distribution procedures has had its impact on the high levels of naval personnel administration. This is evidenced by the fact that an extensive technical program for the classification and distribution of personnel for the post-war Navy has been developed and approved. An enumeration of the salient features of this program will serve to indicate the extent of the impact which classification services have made upon naval personnel administration during the war. Specifically, it is planned that:

1. Enlisted classification and detailing functions at all levels of administration will be integrated by selecting and training Regular Navy officers and enlisted personnel specifically for such duty.
2. Within the Bureau of Naval Personnel a small group of officers and civilian personnel will be set up in the Enlisted Personnel Activity to develop methods, techniques, and procedures designed to maintain and improve the wartime standards of classification performance.
3. At least one general aptitude test and some form of personal adjustment inventory will be administered to all volunteers for enlistment at recruiting stations.
4. A complete classification section will be established as an integral part of the personnel department of all recruit training centers to inform, test, and interview recruits and assign them to the duty for which they are best qualified.
5. At least one qualified officer or enlisted man will be assigned to each service school to counsel and assign trainees to duty and to arrange for the reassignment of school failures.
6. Qualified officers and men will be assigned to all receiving stations and pre-commissioning centers to interview, classify, and recommend proper duty assignments for transient personnel. Such recommendations and detailing will be accomplished in accordance with the *Manual of Navy Job Classifications.*

7. Each ship and shore establishment will be provided with assistance in organizing and operating a classification and utilization program consistent with standard Navy policies.

8. Data which appeared on the Qualifications Card during the war will become an integral part of the man's Service Record. Cross-index files will be established to assist in finding an individual with specific qualifications, but the Service Record will be the only source of complete data for assignment purposes.

9. A revised basic battery of tests will be administered to all recruits and all other regular Navy personnel to serve as a basis for general evaluation of each man's aptitudes, skills, and abilities.

CHAPTER IV

THE PROGRAM FOR TRAINING OFFICER PERSONNEL

AMERICA's traditional substitute for military preparedness is Yankee ingenuity. In peaceful times, the personnel of the armed forces is only sufficient to furnish a basic pattern for the organization and development of the forces which actually fight our wars. At its peak strength during World War II the Navy's officer corps was larger than the entire Navy of 1939, and more than 90 per cent of the officers held reserve or temporary commissions. Obviously the process by which these workers, teachers, technicians, lawyers, students, etc., were transformed into naval officers involved a gigantic program of selection, classification, and training. To a large extent this program itself was planned or improvised and developed by reserve personnel under the guidance and direction of officers of the regular establishment. Implementation of the Navy's policy, "To make effectiveness in war the objective of all development and training" called for the application of both the specialized knowledge of regular officers and the best techniques known to educators and to industrial and personnel management.

Administration of the Training Programs

Cognizance over the training of both officer and enlisted personnel of the Navy is assigned to the Bureau of Naval Personnel.[1] Within the Bureau, principal responsibility for administration of the training programs is assigned to the Training Activity. The wartime organization of this Activity is briefly outlined below.

1. The Standards and Curriculum Division was responsible, in collaboration with field agencies and other agencies within the Navy Department, for determining the training standards required to meet the needs of the service; for the development and promulgation of curricula for all schools and courses; for research in selection procedures, tests, and requirements for ratings; for the development of standardized requirements, tests and procedures, and for the development and supervision of field programs to provide off-duty educational opportunities. To accomplish these ends, the Division was organized into sections designated as

[1] Three exceptions may be noted. The Bureau of Medicine and Surgery maintains control of its professional technical training functions. Training of aviation personnel is under cognizance of the Aviation Training Division of the Office of the Deputy Chief of Naval Operations (Air). Training of personnel assigned to duty with the fleet is under the cognizance of the respective Fleet Training Commands, the Bureau of Naval Personnel providing logistic support.

Billet Analysis, Curriculum, Training Courses, Test and Research, Instructor Training, and Educational Services.

2. The Field Administration Division was responsible for the establishment of training centers and schools as needed; for the development and procurement of the facilities, equipment, and staffs required for the centers and schools; for the continued supervision of the schools and other phases of the training programs, and for the administration of the physical fitness program.

3. The Training Aids Division was responsible for the preparation, evaluation, and distribution of devices and materials designed as aids to training, and for the administration of a program to insure the effective utilization of these materials in the training programs.

4. The Quality Control Division was assigned responsibility for evaluation of the naval training programs in terms of the requirements of the operating forces; for review of the effectiveness of the operations of the training centers and schools, and of the training programs devised and administered by the other divisions of the Training Activity; for liaison with operational commands and with the schools.

Obviously, training had to be coordinated with other activities and operations of the Navy. The manning and equipment of new ships and stations and the maintenance of the allowed complements were basic factors. Predictions of the personnel needs were supplied to the Bureau of Naval Personnel by the Chief of Naval Operations. Quotas of trainees and operating personnel, allowances of equipment and operating budgets for the schools were cleared by the Planning and Control Activity. Selection of trainees for the enlisted service schools was made by classification officers under Enlisted Personnel Activity. Assignment of officers to school staffs was made by Officer Personnel Activity.

Directives originated in the Bureau of Naval Personnel were promulgated via the commandants of the Naval Districts in which the training installations were located. A director of training was assigned to the staff of each commandant to provide technical advice and supervisory services to the training schools located within the district.

Officer Training for the Regular Navy

Officers of the regular Navy undergo intensive training designed to prepare them for the complex responsibilities of command at sea. After graduation from the Naval Academy at Annapolis, their assignments to shipboard duty are rotated so that in the course of a few years they will have served in the various departments in the several types of combatant and auxiliary ships. After several years of this rounding experience, some degree of specialization is provided

through the special fields of the Post Graduate School. The general line course of the Post Graduate School provides a refresher course in the theoretical aspects of naval subjects and additional training for the responsibilities of a commanding officer. The Naval War College offers training for the higher echelons of command. In the periods of duty between the school years, the diligent officer may avail himself of correspondence courses. To some extent, pursuit of such courses is motivated by the prospect of the examinations which the officer must take in order to qualify for promotion to each successive higher rank from lieutenant junior grade to captain.

Obviously, this program is geared to the requirements of the permanent peacetime naval establishment. It is not, nor is it designed to be, a program for the maintenance of a reserve. Alongside of it, there existed before the war a training program for reserve officers. Even as it was set up on paper, this program was inadequate, and the paper program was not realized in practice. As the nation came to a realization of its immediate need for a strong navy and as war became first imminent and then actual, the task of finding and training qualified officers became stupendous.

The solution of the problem lay partly in specialization, partly in the development of streamlined and intensive training programs. In the emergency, officer personnel were drawn from four principal sources: (a) former naval officers recommissioned in the reserve; (b) enlisted men who were given temporary commissions; (c) civilians in administrative positions and in technical and scientific positions related to the specialized needs of the Navy, and (d) college students.

Both the variety of the Navy's requirements for officers and the diversity of the educational and experiential backgrounds from which the officers were procured dictated the development of several training programs. Practically all reserve officers, whether destined for general, technical, or specialized service billets, required some form of naval indoctrination training. Beyond this indoctrination, there were varied needs for training in technical specialties, for specialized types of duty, and for team or group training for different types of ships and operational units. There was also a need for preliminary training by which qualified younger men could be given general preparatory training at the college level.

The officer training programs can thus be divided into three general classes. First, the primary officer schools consisting of the V-7 and reserve midshipmen's schools, indoctrination schools and the college training programs (Naval Reserve Officer Training Corps Units and V-12 Units). Second, advanced officer training schools consisting of the technical and semi-technical schools and the specialized

4073
Indian Institute of Public Administration,
Library, 6, Bhagwandas Road, New Delhi.

training programs. Third, the operational training programs under the cognizance of the fleet commands. A brief description of each of these types of training follows.

Primary Officer Training Schools

V-7 AND THE RESERVE MIDSHIPMEN'S SCHOOLS. The first of the emergency officer training programs to be developed were the V-5 and V-7 programs. V-5 was designed for the training of naval aviation cadets. V-7, as originally authorized in June 1940, accepted officer candidates between the ages of 19 and 26 years, who had completed two years of accredited college work, were unmarried and physically qualified for general service in the grade of ensign. These men were enlisted as apprentice seamen and given thirty days shipboard training, then appointed midshipmen and put in a reserve midshipmen's school for three months of intensive study in seamanship, communications, naval administration, naval engineering, damage control, navigation, recognition, and ordnance and gunnery. Upon completion of this course, they were commissioned ensigns and assigned either to duty or to further instruction.

In 1941, requirements for acceptance in this program were raised so that a college degree with work in mathematics, and age between 21 and 28 were required. After Pearl Harbor there were frequent changes in the specific requirements; married men were accepted; the educational requirement was reduced to the equivalent of two years of college, and to some extent, sea duty was accepted as an equivalent of college work. Enlisted men recommended by their commanding officers were accepted. This led to the establishment of academic refresher units or pre-midshipmen's schools. As the number of trainees and schools increased, a four-weeks "indoctrination" period at the school locations was substituted for the shipboard training. Between July 1940, and the graduation of the last class of reserve midshipmen in December 1945, nearly 70,000 midshipmen were commissioned ensigns of the Naval Reserve. Seven schools were employed. One trained engineering officers only; two others trained both deck and engineering officers; the remaining four trained deck officers.

WAVES officer candidates were selected from civilian life and the enlisted ranks and received primary training in a women's reserve midshipmen's school. The curriculum consisted of basic indoctrination training for all candidates. This was followed, for some officers, by specialized training in communications or advanced indoctrination.

NAVAL INDOCTRINATION SCHOOLS. The indoctrination schools provided intensive training of reserve officers who were commissioned directly from civil life. Procedures of the offices of naval officer procurement are described in Chapter II. In most instances, commissions were accompanied by orders either to active duty or to "active duty under instruction." Officers ordered directly to active duty were required to register for a correspondence course in Navy Regulations. Some activities, to which considerable numbers of newly commissioned officers were assigned for duty, organized part-time indoctrination classes. But the normal process was assignment to "active duty under instruction" at one of the naval indoctrination schools. These schools were established at thirteen different locations and trained nearly 60,000 officers. They provided eight weeks of intensive instruction in naval customs and traditions, navigation, ordnance and gunnery, seamanship, naval regulations, administration and law, communications, and military drill. All of the officers so trained were over 21 years of age and either had college degrees or experience in their special fields which indicated an equivalent level of ability and achievement.

THE COLLEGE TRAINING PROGRAMS. The V-12 college training program, which became the principal feeder into the reserve midshipmen's schools toward the end of the war, was antedated by the V-1 program. Students regularly enrolled as freshmen or sophomores in colleges and universities were enlisted in the grade of apprentice seaman and permitted to continue their college courses in an inactive duty status. At the same time, students in the upper two years of college were enlisted under the V-7 program and held in an inactive duty status. By this means, a pool of young men of officer candidate quality was reserved in college training when the Selective Service age was lowered to 18 and deferments of college students became increasingly limited. The V-1 program was in effect from March, 1942, to July 1, 1943, when the V-12 program superseded it.

The V-12 program was developed on the basis of a joint declaration of policy on the utilization of colleges by the Secretaries of War and Navy. Under this program, young men were selected for college training of from two to eight 16-week terms, depending on the type of service for which they were being prepared. Except for the pre-medical and pre-dental students, the work of the first two terms was uniform. After the first two terms, the candidates were selected for assignment to "upper level specialties" on the basis of quotas of the numbers required, and individual preferences, aptitudes, and ability as indicated by screening tests. The broad outlines of the curricula were set by a Navy Educational Advisory Council, developed into lists of courses by subject matter specialists acting as

consultants, and their work was reviewed by the Navy bureaus and offices concerned. The consultants then prepared brief course descriptions and the whole plan was reviewed and approved by the advisory council. The initial quota of 80,000 trainees was drawn from inactive reservists in the V-1, V-5, and V-7 programs, and Army Enlisted Reserve Corps enrollees who expressed a preference for naval service, from enlisted men selected on the basis of General Classification Test scores and recommendations of their commanding officers, and from civilian high school graduates and college students who took a special examination (described in Chapter VI) and were selected by district boards within the limits of quotas assigned to each state.

V-12 Units were established in 131 colleges, in 73 of the country's 76 medical schools, and in 37 dental schools. The men were enlisted in the grade of apprentice seaman and assigned to active duty under instruction. Quotas were set up for Marine Corps and Coast Guard candidates. Aviation candidates were classed as V-12a, and upon the completion of two terms of the V-12 program, were transferred to class V-5 and put in aviation training. Candidates for deck officer training and general Marine Corps service were transferred to reserve midshipmen's schools and Marine Corps schools after four terms. Pre-medical and pre-dental students were transferred to the professional schools upon completion of five terms. Candidates for Supply Corps and general engineering training were advanced to Supply Corps schools and reserve midshipmen's schools after finishing six terms of the program. Aerology specialists and engineer specialists in the various curricula—naval architecture and marine engineering, civil, mechanical, electrical power, electrical communications, electronics, aeronautical engines, aeronautical structures, and physics—were given eight terms and then advanced to reserve midshipmen's schools. Pre-chaplain candidates were given eight terms and then sent to the seminaries of their denominations for two years or more. After completion of the first two terms, Marine Corps specialist candidates were given courses arranged by the academic authorities for six additional terms in the fields of electrical engineering, electronics, civil, mechanical, and mining engineering, in preparation for Marine Corps schools in the fields of communications, engineering, ordnance, and combat engineering.

Naval Reserve Officer Training Corps Integrated in the V-12 Program. After the establishment of the V-12 program, the Naval Reserve Officer Training Corps Units in 27 colleges were maintained as integral components of the V-12 commands. V-12 students were selected for the Naval Reserve Officer Training Corps during the second term and continued for five additional terms in one of three

standard curricula: general, engineering, or business administration. Upon completion of the Naval Reserve Officer Training Corps curriculum, graduates were commissioned ensigns of the Naval Reserve or second lieutenants of the Marine Corps Reserve. Shortly before the end of hostilities, the authorized strength of the Naval Reserve Officer Training Corps was increased from 7,200 to 24,000, to be reduced to 14,000 one year after the end of the war; 25 additional Naval Reserve Officer Training Corps Units were authorized.

At the same time, liquidation of the V-12 program was planned. Enrollment in the V-12 program had been progressively reduced beginning in November 1944, so that by November 1, 1945, the enrollment was less than half the original allowance. Concurrently, the proportion of students selected from enlisted rather than civilian sources was increased. By November 1945, most of the V-12 students were transferred to Naval Reserve Officer Training Corps Units or were taking naval reserve officer training courses in interim V-12 Units. Ten of the new Naval Reserve Officer Training Corps Units were assigned to institutions which had not had other naval training programs. The curricula prescribed for these units are designed for the normal four-year college course. In the remaining 42 Naval Reserve Officer Training Corps Units, conversion to the four-year program was scheduled to become effective in the summer of 1946.

Advanced Officer Training Schools

TECHNICAL TRAINING SCHOOLS. In the field of technical training, advanced schools were established for officers to be assigned to duty as ordnance, communications, damage control, tactical radar and fighter director officers; as engineering officers on diesel installations; and as officers responsible for the installation and maintenance of electronics gear. Diesel, communications, and electronics training courses were located in civilian training institutions, both colleges and industrial plants. Damage control, ordnance courses, and most other technical and semi-technical programs were developed at naval installations. Most of the officers assigned to these schools for instruction were recent reserve midshipmen's and indoctrination school graduates. A problem that was common to all of the training schools was met in intensified form in the technical schools—the equipment available for instructional purposes was neither adequate nor up-to-date. When the schools were first established, the curricula were developed locally. By September 1943, the curricula were standardized and equipment and educational aids were provided to implement the standardized curricula.

In addition to the technical training of a pre-service nature pro-

vided in these schools, the Post Graduate School of the Naval Academy provided advanced technical training for experienced officers, both regular and reserve, in its naval architecture, engineering, ordnance, communications, radio, law, and staff and command courses.

SPECIALIZED TRAINING COURSES. Specialized training courses were established to meet the Navy's needs for experts in various fields of specialized service. Among the continuing specialized programs were those for oriental languages, military government, chemical warfare, and naval justice. Temporary programs were set up from time to time for specific needs, such as the indoctrination of officers procured from the ranks of secondary school and college teachers and administrators to serve as commanding officers of the V-12 Units.

The oriental language program furnishes an outstanding example of the development of such courses. The outbreak of the war found the country with a conspicuous dearth of men qualified as translators and interpreters of Japanese. A few naval officers had received training in Japanese at the University of Tokyo over a period of nearly twenty years, but of the 60-odd who had taken the 3-year course, only about a dozen were regarded as fully proficient. To meet this need, two schools were set up. Adapting the system of instruction which had been employed at Tokyo, the course was compressed and intensified to a 1-year program. Only men with outstanding linguistic ability and Phi Beta Kappa caliber in general ability were selected for this training. Originally both civilian and service personnel were enrolled, but later selections were made only from officers who applied for the training, or civilian applicants were commissioned and assigned to the schools on a "duty under instruction" basis. In June 1944, courses in Malay, Chinese, and Russian were established.

Operational Training

The types of training described thus far were designed largely to increase the individual officer's knowledge and understanding of the Navy and of the special field in which he was to be assigned to duty. But an effective operational unit requires coordination and teamwork, the development of techniques and routines. Early in the war the need for activities which would build up such teamwork led to a mushroom growth throughout the Navy of training programs intended to overcome the locally discovered deficiencies. The apparent deficiencies resulted both from the relative inexperience of officers and men and from the development of new instruments, methods, and doctrine. Originally created without central control or direction, programs of this kind were gradually integrated into better planned activities partially under the Bureau of Naval Per-

sonnel, partially under the Fleet Operational Training Commands.

The key factor in operational training is the development of efficient team action in the performance of tasks which require varied rather than uniform activities on the part of individuals making up the operational unit. Generally, such efficiency is dependent upon thorough understanding of the unit's task and of the individual's function in the organization, upon the competence of each individual in his job, and upon habitually correct reaction to the problem situations occurring. In operational training, therefore, the trainees are organized into groups that work together, a gun crew, a repair party, or a combat information center team. Here newly commissioned officers may experience their first command contact with enlisted men and their first responsibility for successful completion of a task. Particularly in the pre-commissioning schools, they learned in practical and realistic terms, under the compelling motive of having immediate need to use all they could learn, the functions and responsibilities of the billets to which they were assigned.

Approximately fifty different kinds of operational training schools were set up. The following list is only illustrative of the variety: Advanced Base Supply Training, Advanced Base Aircraft Carrier Gasoline Systems, Amphibious Training, Armed Guard, Combat Information Center, Team Training, Harbor Defense, Mine Disposal, Submarine School, Destroyer School, Salvage, Wartime Merchant Ship Communications.

Post-War Training Plans

The program of officer training for the post-war Navy depends upon the action of Congress. Plans currently under consideration propose to broaden the basis for selection of officer candidates and officers. Under legislation now in effect, college graduates with Naval Reserve Officer Training Corps training are eligible to transfer to the regular Navy after a period of active duty and examination. The allowed peacetime enrollment in the 52 authorized Naval Reserve Officer Training Corps units is 14,000 men, which should produce between 2,500 and 3,000 graduates per year. The amount of training in Naval Science subjects which will be required in these units has not been determined and may depend to some extent upon the provisions made for financing the program. It is also proposed that graduates of accredited four-year colleges without Naval Reserve Officer Training Corps training be made eligible for commissions as officers in the Reserve with similar privileges of appointment to the regular Navy after experience and examination.

These plans are formulated in terms of supplying the needs for

officers of the projected post-war Navy. As this is being written, plans for maintaining a progressive training program for reserve officers are less well developed. Should the need arise for maximum expansion of the Navy, it is to be expected that many of the expedients adopted during World War II will again prove useful, and further that facilities, personnel, and equipment for officer training programs will be recognized as needs equally urgent with the need for new ships. The experience of World War II has proven conclusively that the complex gear aboard modern fighting ships can be used effectively only by highly trained men under competent officers.

CHAPTER V

THE PROGRAM FOR TRAINING ENLISTED PERSONNEL

DURING the war the Navy found it necessary to man its enlisted billets predominantly by civilians who had neither experience in nor training for the military and highly technical work which had to be performed aboard ship and at shore stations. Training these individuals for their jobs became one of the most critical undertakings of the Navy. Production of trained personnel needed to keep pace with construction of new ships and aircraft. Consequently, personnel had to be trained by the hundreds of thousands in relatively short but highly intensified training programs. Then, too, as the war progressed, under the impact of new inventions and improvement in fighting equipment, skills and knowledge already acquired through training frequently became obsolete. Under these circumstances, the Navy training program was subject not only to tremendous expansion, but also to continuous revision, both in terms of length and content of specific courses, and with respect to the number of personnel in various training programs.

Scope of the Enlisted Training Program

The structure of the training program for enlisted personnel had already been established before the outbreak of the war. It involved:

1. an introductory period of basic training for recruits (given in Naval Training Stations prior to May 1944, thereafter in Recruit Training Commands in Naval Training Centers);
2. a system of Naval Training Schools to which selected personnel were assigned for elementary and advanced technical training in the various enlisted specialties which are distributed among eight branches of service: Seaman, Artificer, Artificer-Engine Room Force, Aviation, Special, Commissary, Specialist, and Steward;
3. in-service training programs for advancement of personnel in their specialties.

This structure of training was adapted to wartime needs by expanding the Navy school system, by accelerating training in all its aspects, and by developing special programs to meet new demands and temporary needs. During the period from 1939 to 1945, approximately 4,000,000 enlisted men and women were trained in one or more of these programs.

Specialization and Promotion of Enlisted Personnel

Two principal motivating factors behind Navy training have been (1) a system of rates (or ranks) which enabled personnel who had acquired technical training in a specialty to be promoted or advanced on the basis of increased proficiency, and (2) a scale providing for substantial increases in the base pay of personnel upon their advancement from one rate or pay grade to another.

TABLE 1-v. Base Pay for Navy Rates

RATE	BASE PAY PER MONTH *
Apprentice Seaman	$50.00
Seaman or Fireman, second class	54.00
Seaman or Fireman, first class	66.00
Petty Officer, third class	78.00
Petty Officer, second class	96.00
Petty Officer, first class	114.00
Chief Petty Officer (acting)	126.00
Chief Petty Officer (permanent)	138.00

* These rates prevailed on August 15, 1946.

As shown in Table 1-v, this structure enables individuals to advance through a series of steps from Apprentice Seaman, the level at which practically all enlisted personnel enter the Navy, to Chief Petty Officer, a rate which is attained by a relatively small per cent of men. Since increased technical or special skills are required for each advancement in rate, and consequent increase in base pay, most personnel are motivated to obtain the training which would help them qualify for advancement. This training is obtained in Navy schools and in local apprenticeship programs aboard ship and at shore stations.

Training Schools

Training schools of three different types were maintained to supply personnel for both general and technical duties ashore and afloat. The first of these were the establishments for recruit training, to which all new personnel were assigned upon enlistment or induction, to acquaint them with the general duties of enlisted naval personnel. Training of the second type was given in service schools to those individuals who (in terms of scores on basic battery tests) appeared to be best qualified for technical training in some one of the many enlisted specialties. A third and more advanced type of training, of an operational character, was provided for personnel whose duties would involve working in a team or group.

RECRUIT TRAINING. Upon entering naval service, newly enlisted or inducted personnel were first ordered as apprentice seamen to active duty for a period of recruit training. During the war this training program varied in length from three to twelve weeks and was given in one of seven naval training stations or recruit training commands. In contrast to later training, the instruction on the recruit level was less technical and was concerned in the main with giving recruits an overall view of the Navy, its mission, its operations, and the manner in which each individual was expected to contribute toward the accomplishment of that mission. Instruction was given in elementary seamanship, fire fighting, first aid and personal hygiene, military drill, physical conditioning and swimming, lookout-recognition, gas warfare defense, use of sound-powered telephones, and ordnance and gunnery.

During the early part of his recruit training, each individual was tested with the Basic Test Battery of classification tests (see Chapter III for a complete description of the classification program). Performance on these tests was motivated by informing the recruits of the types of billets to which they might be assigned, and by explaining that specific assignments would be conditioned in part by special abilities and aptitudes indicated by test scores. A few days after the administration of the tests, an interview was scheduled in which the interviewer obtained relevant information regarding civilian work experience, education, and interests of each individual. Such information was used along with test scores in determining whether the individual should, upon the completion of recruit training, be recommended for general detail aboard ship or at a shore station or whether he could be of most service to the Navy if given specialized technical training for one of the highly technical jobs. Assignments were based upon the recommendations of the interviewers and upon changing needs for distributing the Navy's manpower.

Recruit personnel were advanced automatically to seaman or fireman, second class, at the end of basic training. Thereafter, advancements within an enlisted specialty were made individually in accordance with requirements in each general service rate. If assigned to general duty, the recruit was encouraged to select an enlisted specialty and to obtain in-service training which would help him advance in his rating. If assigned to an elementary service school. the training received there would enable the recruit to qualify more rapidly for advancement in rating upon his assignment to a billet.

Standardized written achievement examinations produced in the Test and Research Section were administered routinely to all recruits at the completion of their training. (See Chapters XIV and XV). Since these examinations were based on the informational content

of the curriculum, trends in test results were used as criteria in evaluating the effectiveness of instruction in the recruit training program.

ELEMENTARY SERVICE SCHOOLS. Two types of elementary service schools were maintained to provide training on a pre-rating level for quotas of qualified trainees drawn from the recruit population. These were "P" schools and "A" schools.

"P" schools were maintained (1) to provide short-course training in elementary operating skills needed by personnel who were to be assigned to certain basic types of duty, and (2) to provide training which would prepare selected personnel for a somewhat more advanced type of technical training in "A" schools. The period of instruction in "P" schools varied from three to eight weeks. Graduates of these schools were either assigned to further training or to duty in the enlisted specialty for which they were trained.

"A" schools were maintained to provide the elementary technical training needed by third class petty officers in various specialties. The period of instruction in "A" schools varied from six to twenty weeks. In general, however, training in most "A" schools lasted for sixteen weeks. In the earlier months of the war, a certain percentage of the best qualified graduates of "A" schools were advanced to third class petty officer ratings at the end of their training. The remainder were graduated as non-rated seamen or firemen, first or second class, and given designators to facilitate their assignment to the duty for which they were trained. While the practice of rating graduates of "A" schools was later discontinued, their advancement in rate was more likely to follow in duty billets as a result of the specialized training obtained.

The Test and Research Section was responsible during the war for constructing standardized achievement examinations and performance tests for use in both "A" and "P" schools. A description of examinations and outcomes of testing in these schools will be found in Chapter XV.

ADVANCED SERVICE SCHOOLS. Included in this classification are "B" and "C" schools. While both of these types of schools drew trainees from the same sources, namely, elementary schools, shore centers, and the fleet, their curricula differed widely.

The curricula of "B" schools were designed to prepare trainees to meet the technical qualifications required for a general service rate of first class petty officer. Trainees assigned to these schools were second and third class petty officers and the period of training varied in different schools from eight to fifty-two weeks. Typical of the "B" schools were those for torpedomen, aviation metalsmiths, and fire controlmen.

"C" schools were designated by type as "C-1" schools, located in naval training centers, and "C-2" schools, operating in manufacturing plants and factories. Their curricula were designed to provide training in the development of special skills and knowledge required in specific types of duty in general service ratings or in maintaining and operating new equipment. For example, among the "C-1" schools maintained during the war were those for training printers, typewriter repairmen, and motion picture operators. There were "C-2" schools for such training as that needed by teletype maintenancemen, diesel instructors, and aircraft instrument repairmen.

Fleet Service Schools and Fleet Schools. Both of these types of schools were maintained to supply the fleet with personnel equipped with special abilities needed only in shipboard billets. Curricula were therefore shaped in nature and scope to meet specialized manpower needs of the fleet. For example, among the elementary and advanced types of training provided in these schools was that needed by landing craft operators, ammunition handlers, welders, communications technicians, and aircraft maintenancemen. The main difference between these two types of schools was one of administrative control. While Fleet Service Schools were operated for the fleet by the Bureau of Naval Personnel, all Fleet Schools were under the cognizance of Fleet Training Commands.

Operational Training. Operational schools were maintained for the purpose of providing team training for groups. Personnel in operational training often included both officers and enlisted men who were assigned to this training from the fleet, from recruit training, or from Navy schools. These trainees were organized for instructional purposes into tactical units similar to later organizations of units ashore or aboard ship. For example, this type of training was given the nucleus crews of new construction. Instructional emphasis was placed in operational training on the development and coordination of skills needed in operating such shipboard equipment as that used by groups in fire fighting, salvaging and repair, handling ammunition, and in bomb disposal.

Training for Advancement in Rate

Training in Navy schools and local apprenticeship programs of training for advancement in rate were maintained to supply skilled personnel for petty officer ratings. While training in a Navy school did not in itself qualify an individual for advancement to petty officer status, the training he received there usually enabled him to meet the requirements for advancement soon after he had been assigned to a billet in his specialty. But all personnel, regardless of

previous training or the lack of it, were eligible for training for advancement. This training could be obtained in each activity or afloat by "striking" for the next higher rate in an enlisted rating and learning while working in an assigned duty billet.

In general, a non-rated man or a petty officer became eligible for advancement to the next higher rate in his specialty when he had (1) spent a specified length of time in his present rate or billet, (2) completed a Navy training course prescribed for the specialty and the rate for which he was "striking", (3) attained at least the minimum marks specified in proficiency and personal conduct, (4) demonstrated proficiency in the technical skills of the rate for which he was "striking", and (5) passed a technical examination based on the training course issued for his study. Advancement within enlisted specialties to petty officer status was therefore contingent upon meeting requirements which pertained to the training, length of service, and personal attributes of personnel. Since these requirements were prescribed by the Bureau of Naval Personnel, procedures in training for advancement were uniform in naval activities, and relatively high standards of proficiency were maintained.

Programs of training for advancement in rate were organized along similar lines in most activities. A training officer was placed in charge of the local program. Responsibilities for supervising the progress of personnel in training were delegated to division officers, and within divisions, to senior petty officers.

In conducting a local in-service training program for advancement in rate, general procedures were as follows: (1) personnel were acquainted with the rates to which they were eligible to advance and with the specific requirements for advancement, (2) training course manuals were issued by training officers, (3) opportunities were provided for personnel to qualify in the practical phases of the work involved, (4) progress tests were administered either at the end of a training course or from time to time during the course as a means of determining how well prepared trainees were, (5) a final technical examination was prepared by an examining board and administered in order to determine the technical competence of personnel who had completed a training course, and (6) those who passed the final examination to the satisfaction of the examining board were recommended by the commanding officer for advancement to the next higher rate.

Special Training Programs

In addition to the major types of training described in the preceding section, programs of lesser scope were organized to meet

special needs for limited periods of time. The more important of these special training programs included (1) Women's Reserve Training, (2) Refresher Training, (3) Organized Shipboard Training, (4) the Retraining and Disciplinary Program, and (5) the Special Recruit Training Program.

WOMEN'S RESERVE TRAINING. Since WAVES were procured to replace men in shore billets, the program of training enlisted women was similar to that of enlisted men. Following a period of six weeks in recruit training, WAVES personnel were assigned, depending on their qualifications, to general duty or to special training.

Special training programs in Navy schools were maintained for enlisted women to prepare them for duty as yeomen, storekeepers, mailmen, gunnery trainer instructors, hospital corpsmen, and for certain aviation and radio ratings. If assigned to general detail, WAVES personnel could advance in any of the ratings open to women by "striking" for a given rating and qualifying for each advancement.

REFRESHER TRAINING. Programs of refresher training were designed specifically to meet the temporary training needs of a ship's crew while the ship was in port and waiting for repairs, fuel, or supplies. Curricula were necessarily flexible since they had to be adapted to the length of time available for training. Instruction was given in both elementary and technical subjects. Thus, refresher training was given to maintain or increase skills which might otherwise have deteriorated over periods of inactivity.

ORGANIZED SHIPBOARD TRAINING. Shipboard training had as its purpose the improvement of the individual as well as group performance of personnel. An officer on each ship was assigned to administer the program; training included organized programs of team training, refresher training, and training for advancement in rate.

RETRAINING AND DISCIPLINARY PROGRAM. The retraining and disciplinary program during the war was organized as a means of restoring offenders of military law to naval duty or, if discharged, to rehabilitate them for civilian living. To this end, a constructive retraining and vocational training program was maintained for prisoners who had been sentenced to varying periods of confinement by general courts martial.

Each prisoner was first placed in quarantine for a period of a few days. During this time, he was interviewed and various psychological and educational tests were administered. The offender was then brought before an assignment board which in the light of his mental, moral, and health status, recommended his assignment during confinement to some type of vocational training.

Vocational training occupied the major portion of a prisoner's

period of confinement. It involved work programs in which specific items of naval equipment were produced or in which equipment was salvaged and repaired.

While vocational training was given all military offenders under sentence by a general court martial, the retraining program was designed specifically to help personnel readjust to the demands of active naval duty. Accordingly, prisoners who were to be returned to duty were assigned from vocational training to ten weeks' retraining immediately preceding their release from confinement.

Upon assignment to the retraining program, personnel participated in physical training and military drill, and were given a refresher course in naval orientation. In addition, training in an enlisted specialty was provided in basic seamanship, gunnery, communications, pre-engineering, or in yeomen, steward's mate, and ship's cooks and baker ratings.

Special Recruit Training Program. Beginning in June 1943, it became necessary to accept illiterates for naval service. These illiterates should not be confused with men rejected from service for reasons of low mentality. They included non-English speaking inductees as well as men who for certain environmental reasons lacked basic skills in reading, writing, and arithmetic. Technically, an English speaking illiterate was one who had at least sufficient intelligence to be useful in the service, but was below a fourth grade level of proficiency in reading, writing, and arithmetic skills.

It was obvious that optimum use could not be made of these individuals until they had developed at least minimum proficiency in reading, writing, and use of numbers. It was equally obvious that the existing recruit training program was not adequate for the needs of these illiterates. Consequently, a special training program of 14-weeks' duration was organized for these handicapped recruits. While the most important single objective in this special training program was that of providing literacy training, provision was also made for military and physical training, and for instruction in elementary seamanship and related topics.

It was necessary to organize each company for instructional purposes into three groups. The first group included the totally illiterate. The second group was made up of men who approached a fourth-grade level of skills in reading, writing, and arithmetic. The third group consisted of non-English speaking trainees. Initial classification of trainees into these three groups was made on the basis of performance on the Navy Literacy Test, the Navy Non-Verbal Classification Test, (see Chapter VI) and the Special Training Writing Test.

Instructors, who for the most part had been teachers in civilian

life, were carefully chosen and given in-service training in handling special problems which were likely to arise in this program. It was important, for example, to correlate instruction in reading, writing, and arithmetic with military and physical training and with seamanship. Instruction had to be adapted to the language and comprehension level of trainees. Emphasis upon learning by doing was made through wide use of visual aids, demonstrations, and classroom discussions.

Standardized forms of achievement examinations in reading were produced in the Test and Research Section for use in this special training program. These examinations were used along with the progress tests and other achievement examination materials to evaluate the instructional outcomes of special training.

Post-War Training

Plans for the post-war training of enlisted personnel are still in the process of being developed. To a large extent, these plans must await action of Congress on the size of the Navy. In the main, however, no large scale reorganization of enlisted training is anticipated.

The basic pattern of recruit training, followed by technical training in Class "A", "B", and "C" schools, will continue in the post-war training program. In addition, it is tentatively planned that qualified recruits will be sent to primary schools for an initial period of general technical training prior to shipboard assignment. Following a tour of sea duty, selected personnel will be assigned to Class "A" schools for training in the various specialties. Advanced technical training will continue to be maintained in the Class "B" and "C" schools. Finally, training programs for advancement in rate will continue to serve all enlisted personnel who are striking for ratings within their specialties.

PART II

THE CONSTRUCTION, STANDARDIZATION, AND USE OF SELECTION AND CLASSIFICATION TESTS



CHAPTER VI

BASIC TESTS FOR ENLISTED PERSONNEL

In 1942 a battery of mental tests which had been in use for several years was employed in most of the recruit training programs of the United States Navy for the purpose of classifying enlisted personnel. The battery consisted of the following tests:

1. O'Rourke General Classification Test—Junior Grade (U. S. Navy Edition).
2. Mechanical Aptitude Test—Junior Grade (U. S. Navy Edition).
3. Standard Test in Arithmetic.
4. Standard Test in Spelling.
5. Standard Test in English.
6. Radio Code Aptitude Test.

These six tests were used to classify recruits as to their aptitudes for specialized training in naval training schools.

Extended study of these tests led to the conclusion that they failed in several respects to fulfill their functions adequately. Studies were initiated in 1942 to determine the values and weaknesses of the tests and to appraise the procedures used in their administration and scoring.[1] At the same time the decision was made to proceed with the construction of a new battery of tests.

A study of the accuracy of scoring the tests at four naval training centers showed that large errors were commonly made, some great enough to render the test scores valueless for selection purposes. Steps were taken to eliminate the major sources of error by devising a new standardized scoring procedure which was strongly recommended to all training centers.

The results of the studies of the validity of the old basic test battery for predicting success in service schools fully confirmed the earlier judgment that new tests were desirable. For example, a study was made of the validity of the old tests in six different service schools at a large naval installation. It was found that the correlation coefficients between test scores and service school grades were so low that the tests were of questionable utility. A study of an experimental battery of tests at three advanced service schools showed that the most valid of the experimental tests contained items similar in type to those in the proposed new test battery. Another analysis of

[1] This chapter is based upon research reports prepared by the NDRC Project N-106 and the College Entrance Examination Board and upon unpublished studies of the Test and Research Section, Bureau of Naval Personnel (see Appendix C).

the old battery of tests showed that some of them were too easy, while others were too difficult for optimal use. In particular, the General Classification Test contained too many easy items and the Radio Code Aptitude Test too many difficult ones. The distributions of total scores were generally either skewed or excessively flat, which indicated that the difficulty-values of the items were not well suited to the groups tested.

The findings of many of these studies confirmed the wisdom of the decision made earlier that a new and more effective battery of tests was required for the purpose of classifying large numbers of men with respect to their aptitude for specialized training.

Development of a New Basic Test Battery

It was decided first that the new battery should include tests of verbal, mechanical, and quantitative reasoning abilities. Factorial analyses have shown these abilities to be reasonably well-defined and relatively independent. In addition, many of the billets for which enlisted men would be trained require one or more of these abilities.

At the time the development of the new Basic Test Battery was under consideration, there was a great increase in the number and kinds of jobs which required mechanical information and skill. It was thought, therefore, that tests of reading ability and of mechanical and electrical knowledge would have potential value in the battery. The Reading Test was to measure the candidate's ability to acquire information from texts, while the tests of mechanical and electrical knowledge were to measure the amount of such information the candidate had acquired in his experience.

Tests of spelling, English, and radio code aptitude were thought to be of less immediate value and were therefore not included. The Basic Test Battery, as outlined at the time the development of the tests was begun, was to include the following:

1. General Classification Test.
2. Reading Test.
3. Arithmetical Reasoning Test.
4. Mechanical Aptitude Test.
5. Mechanical Knowledge Test (Mechanical Score).
6. Mechanical Knowledge Test (Electrical Score).

The method employed, as far as possible, in the development of the Basic Test Battery was to construct an experimental form of each test containing a considerably larger number of items than was needed for the final form. The experimental forms were ad-

ministered to sample groups chosen to be representative of the total recruit population and large enough to permit reliable item analyses to be made.

The experimental forms of three tests in the new Basic Test Battery were developed and administered by the end of March 1943 to obtain data for item analyses. The tests were administered in six naval training stations to insure that the sample would contain no geographical bias. Of the total national sample, the group tested at each station was proportional to the total intake at that station over a long period of time. On the basis of a statistical analysis of the results of these experimental administrations, the tests were revised by retaining only those items which met the criteria of validity and difficulty adopted for inclusion in the final test forms. Other items were improved and still others were discarded. The tests were then printed in booklets and were ready for routine administration in June 1943. Owing to pressure of time, the other tests were issued in Form 1 without item analysis of experimental forms.

Three forms of these tests have been widely used, and two revised forms were ready for use shortly after the termination of hostilities. After Form 1 had been put into national use, new tests of clerical aptitude, spelling, and radio code aptitude were added to the battery. A Fleet Edition of six tests of the battery has also been prepared, including the General Classification Test, Arithmetical Reasoning Test, Mechanical Aptitude Test, Mechanical Knowledge Test, Electrical Knowledge Test, and Clerical Aptitude Test.

Studies of the first two forms of the Basic Test Battery showed that some of the tests would profit from revision, particularly the Mechanical Knowledge Test. These revisions were made and incorporated into two new experimental forms, X-4 and X-5, the first of which was being administered experimentally when hostilities ceased and many service schools closed. As a result, progress in the development of Forms 4 and 5 was impeded. The data presented in this chapter are based on the extensive use of Forms 1, 2 and 3.

The tests of the Basic Test Battery were designed so that all items were of multiple-choice form, except for two subtests of the Clerical Aptitude Test. Separate answer sheets were supplied with all of the tests except for the Fleet Edition. The use of separate answer sheets adapted the tests to machine-scoring and at the same time made possible repeated use of the test booklets.

The general tests of the battery were bound in four test booklets as follows:

Book 1. General Classification Test.
Book 2. Reading Test and Arithmetical Reasoning Test.
Book 3. Mechanical Aptitude Test.

Book 4. Mechanical Knowledge Test (Mechanical and Electrical Scores).

The special aptitude tests were issued as follows:

Book 1. Clerical Aptitude Test and Spelling Test.
Record Album. Radio Code Test—Speed of Response.

The Fleet Edition was issued in three booklets as follows:

Book 1. Electrical Knowledge Test, Mechanical Knowledge Test (minus verbal items), General Classification Test (minus opposites items).
Book 2. Mechanical Aptitude Test and Arithmetical Reasoning Test.
Book 3. Clerical Aptitude Test.

General Tests

GENERAL CLASSIFICATION TEST. The General Classification Test was set up to include only items involving verbal ability. Three types of item were included: Sentence Completion, Opposites, and Analogies, in order to require thought and reflection of the inductees, rather than vocabulary or special information exclusively. Such items usually form a test of high reliability.

Illustrations of the three types of item included in the General Classification Test follow:

Sentence Completion

A good sailor will ..... the orders of his superior officers.
(1) see (2) fear (3) read (4) obey (5) like

Opposites

GRIEF
(1) anger (2) poverty (3) joy (4) sorrow (5) pride

Analogies

WATER is to *sponge* as INK is to .....
(1) pen (2) bottle (3) write (4) blotter (5) desk

READING TEST. The Reading Test consists of six paragraphs of increasing difficulty, each followed by several questions. The questions were designed to measure the ability to note details in the material read, to draw inferences from it, and to follow directions. The paragraphs all consist of material related to Navy life. The following is an example of the type of item used in this test:

> After a can of paint has been opened and the paint partly used, the can should be covered and kept as airtight as possible to pre-

vent a paint scum from forming on the surface. If scum forms, the paint should be strained through a fine-mesh wire screen or cheesecloth.

To prevent scum from forming in a partly-used can of paint, one should

(a) keep the can free from dirt.
(b) fill the can up with water.
(c) stir the paint well before storing.
(d) keep the can tightly covered.
(e) make sure that the can is more than half full.

The method of removing scum from paint is

(a) to stir the paint thoroughly.
(b) to pour the scum off.
(c) to strain the paint through a wire screen.
(d) to skim the scum off with a putty knife.
(e) not discussed in this paragraph.

The method of thinning paint is

(a) to mix it with turpentine.
(b) to mix it with linseed oil.
(c) to mix it with white lead.
(d) to mix it with water.
(e) not discussed in this paragraph.

ARITHMETICAL REASONING TEST. The Arithmetical Reasoning Test consists of verbally stated problems which involve ability in quantitative thinking. It was attempted to make the statement of each problem sufficiently simple that reading comprehension would not be an important factor. The actual reading time was intended to consume only a small portion of the testing time. The problems were chosen so that they could be solved by arithmetic, with the computational processes simple enough to require only a small part of the testing time. In other words, the items were chosen and stated so that more than half of the testing time would be devoted to determining the method of solution. The following illustrates the type of item used:

During World War I, keels were laid at Hog Island at the rate of one every five days. At that rate, how many keels were laid in 275 days?

(a) 51 (b) 55 (c) 75 (d) 270 (e) 1375

MECHANICAL APTITUDE TEST. The Mechanical Aptitude Test consists of three sections: Block Counting, Mechanical Comprehension,

and Surface Development. All three types of items involve the ability to perceive visually the mechanical details of a situation which is shown pictorially or graphically, to follow directions, and, in the case of Mechanical Comprehension, to apply to the solution of the problem the physical principle which governs the situation. The spatial abilities involved in Block Counting and Surface Development, in addition to being heavily involved in mechanical aptitude, were considered important in billets in which men would be required to read drawings and blueprints, or to work on patterns.

The types of item used in the Mechanical Aptitude Test are as follows:

Block Counting. This test consists of 2 columns of blocks of the same size and shape as shown in Figure 1-VI. Some of the blocks are

SAMPLE COLUMN

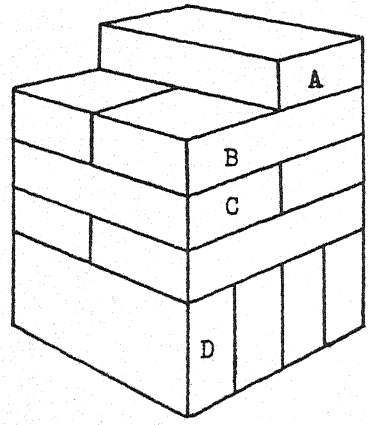


Figure 1-VI. Sample block counting item.

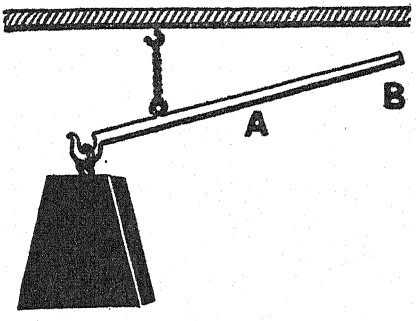
lettered alphabetically. The task is to decide how many blocks touch each of the lettered blocks.

Mechanical Comprehension. In this subtest each item consists of a picture which illustrates a mechanical situation about which a question is asked. Figure 2-VI illustrates an item of this type.

Surface Development. In this test each item consists of a drawing of a flat pattern which can be folded along indicated lines, and of a perspective drawing of the solid object formed by folding the pattern as indicated. In the drawings of the pattern and of the object one side is lettered. A number of lines in the pattern are numbered, while in the perspective drawing of the object all of the edges and some of the sides are lettered. The task is to match the numbered lines of the pattern with the lettered lines of the picture. This task

involves the ability to perceive the relationships involved in transforming from a two-dimensional projection to three dimensions as shown in perspective. An illustrative item is shown in Figure 3-VI.

MECHANICAL KNOWLEDGE TEST. This test contains items of electrical and mechanical knowledge. Part I, called Tool Relationships, consists of pictorial items of which 25 are electrical and 35 are mechanical. Part II of this test contains 80 information items, each

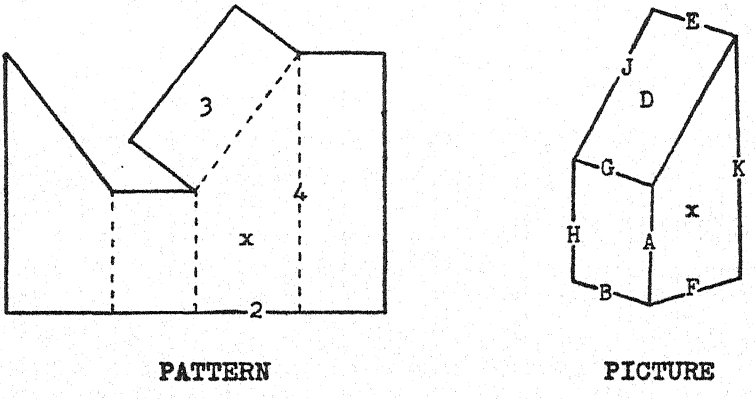


At which point should one pull down to raise the weight more easily?

(1) One should pull down at point A.
(2) One should pull down at point B.
(3) The weight may be raised just as easily from either point A or point B.

Figure 2-VI. Sample mechanical comprehension item.

SIDE (x) IN THE PICTURE IS MARKED IN THE PATTERN.



PATTERN PICTURE

Figure 3-VI. Sample surface development item.

stated in verbal form as a question with which four alternative answers are given. The task is to indicate the correct answer. Two pictorial items are shown in Figure 4-VI, Sample A being electrical and B mechanical. Below each illustration is an electrical and a mechanical information item.

Special Tests

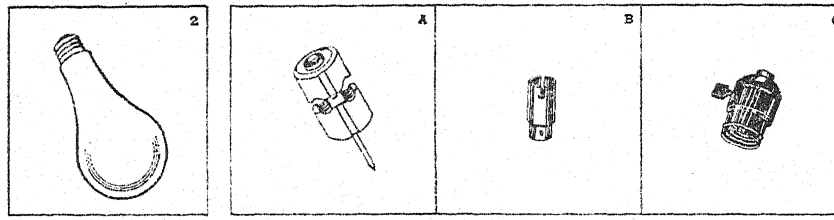
Three special aptitude tests were added to the general tests to complete the battery. These were a Clerical Aptitude Test, a Spelling

Test, and a Radio Code Test—Speed of Response.[2] These tests were designed to measure special aptitudes considered important for success in several types of training schools.

CLERICAL APTITUDE TEST. This test consists of three parts: Alphabetizing, Name Checking, and Number Checking. Aptitude along these lines would appear to be required in storekeeper or yeoman billets.

Alphabetizing. This subtest measures the ability to arrange words in alphabetical order. Each item consists of five words, four of which are given in alphabetical order. The fifth, the key word, is to be

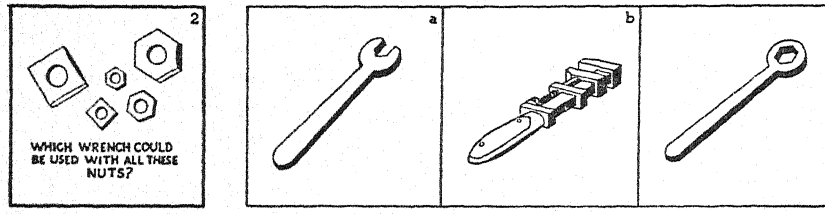
Pictorial Item (Electrical).



Electrical Information Item.

In an a-c operated radio receiver direct current is needed for the
(a) rectifiers (b) plate voltage (c) power pack (d) field coils

Pictorial Item (Mechanical).



Mechanical Information Item.

A rasp is a kind of
(a) saw (b) wrench (c) file (d) hammer.

Figure 4-VI. Sample tool relationship items.

inserted into the correct position so that all five will be in alphabetical order.

CAT____ apple____ barn____ dog____ rat____
1 2 3 4 5

Name Checking. Each item in this subtest, which is designed to measure an aptitude important in clerical work, consists of two names which are either identical or very similar. The task is to state whether the two names in each pair are the same or different. The

[2] The Radio Code Test—Speed of Response is described in Chapter VIII.

differences between members of a pair of names might involve one letter of a name, a change in punctuation, or a grosser and more perceptible difference. The following items illustrate this subtest:

1. J. H. Hornblow & Co.—J. H. Hornaday & Co.
2. American Outdoor Adv'g.—American Outdoor Adv'g.
3. National Life Ins.—National Life Ins
4. Colgates' Dental Cream—Colgates Dental Cream

Number Checking. Each item in this subtest consists of two numbers of from five to eleven digits each. The task is to indicate whether the two numbers in each item are the same or different. The following are sample items:

1. 436815——435815
2. 9614328——9614328

Spelling Test. Each item of this test consists of five words, one of which is misspelled. The task is to indicate the misspelled word, as in the following list:

(1) literature (2) accomadate (3) matrimonial (4) prerogative (5) ellipse

Radio Code Test—Speed of Response. This test, designed to measure aptitude for code learning, is described in Chapter VIII.

Uses of Basic Test Battery

The primary use of the Basic Test Battery has been for classification of enlisted men for Navy training schools. The tests have been administered to all recruits at naval training centers. On the basis of test scores and other considerations, the decision has been made as to whether or not an individual should be recommended to a school for special technical training.

Each test of the battery has demonstrated its usefulness in this selection process, but some of the tests have been more effective than others in selecting men for a specific type of school. For example, for basic engineering and electrical schools, where the curricula require a good deal of mathematical work, the Arithmetical Reasoning Test has been most effective in predicting success. For diesel, torpedoman, and machinist's mate schools, the tests of mechanical aptitude and mechanical knowledge have been most valid. Cutting scores on one or more of the tests have been established as acceptance requirements for 46 types of enlisted men's training programs.

A second use of the Basic Test Battery has been to provide a measure of the quality of the enlisted men and women who enter the Navy. The test scores of all recruits in each naval training center

have been forwarded bi-weekly to the Test and Research Section, Bureau of Naval Personnel. Statistical summaries have been prepared each month showing the mean, standard deviation, and range of scores on each test in each center and in the total recruit population. These data have been used mainly for two purposes: (1) to compare the recruits at different centers over the same period of time, and (2) to study the variation in test scores at all centers over the course of time.

DESCRIPTION OF BASIC TEST BATTERY

Name of Test	Number of Items		Time Limit in Minutes	
	Total	Part	Total	Part
GENERAL CLASSIFICATION TEST	100		33	
Part 1 Sentence Completion		30		10
Part 2 Opposites		30		8
Part 3 Analogies		40		15
READING TEST	30		25	
ARITHMETICAL REASONING TEST	30		30	
MECHANICAL APTITUDE TEST	129		34	
Part 1 Block Counting		45		6
Part 2 Mechanical Comprehension		44		20
Part 3 Surface Development		40		8
MECHANICAL KNOWLEDGE TEST	135		37	
Part 1 Tool Relationships				
Electrical		25		6
Mechanical		30		6
Part 2 Information				
Electrical		35		12
Mechanical		45		13
The following raw scores are commonly derived from the test:				
a. Electrical Score (60 items), Tool Relationships plus Information.				
b. Mechanical Score (75 items), Tool Relationships plus Information.				
CLERICAL APTITUDE TEST	213		13	
Part 1 Alphabetizing		55		4
Part 2 Name Checking		83		5
Part 3 Number Checking		75		4
SPELLING TEST	50		12	
RADIO CODE TEST	150		30	

BASIC TEST BATTERY (FLEET EDITION)

Name and Description of Test	Number of Items in Each Test[3]	Suggested Time Limit in Minutes[4] For Each Book	For Each Test
BOOK ONE		55	
Directions			5
Electrical Knowledge Test	45		10
Mechanical Knowledge Test (minus verbal items)	45		10
General Classification Test (minus the opposites items)	70		30
BOOK TWO		60	
Directions			5
Mechanical Aptitude Test	50		20
Arithmetical Reasoning Test	30		35
BOOK THREE		10	
Clerical Aptitude Test	250		10

[3] The items of each test are of the same type as in the corresponding test of the Basic Test Battery.

[4] The time limits for the separate tests of Books One and Two are suggested time limits only. The subjects are timed on each Book as a whole.

A third use of the Basic Test Battery has been at advanced classification centers and aboard ship, where scores on one or more of the tests have been taken into consideration in the assignment of men to stations or billets. For example, all candidates for the submarine service are required to have a score of 50 or higher on the General Classification Test. Requirements in terms of scores on the tests of the basic battery have similarly been set up for other jobs.[5]

Analyses of the Basic Test Battery

The Basic Test Battery has been the subject of a number of studies, some of which were conducted during the development of the tests in the interest of perfecting them as instruments of measurement. Other studies were conducted later in order to determine the internal characteristics of the tests, such as reliability, and their relationships with external factors, e.g., their validity. In the first category of studies, conducted during the process of developing the tests, were item analyses and time-limit studies.

[5] The program of selection and classification for which the Basic Test Battery was used is discussed more completely in Chapter III.

Developmental Studies

ITEM ANALYSES. With the exception of Form 1, which had to be put into general use without preliminary experimental analyses, the procedure employed in developing the tests was to prepare preliminary forms which contained an excess of items and to administer these to sample groups which were chosen to be representative of the total recruit population. Analyses were then made of all of the items in the preliminary tests, and on the basis of the results of the analyses the best items were retained for the final forms of each test, which were used nationally.

In the item analyses of the Form 1 tests, the primary emphasis was placed on the correlation of each item with the subtest of which it was a part. For each item, the mean total score on the subtest was computed for those persons who selected each alternative answer and for those who omitted the item. The value of each item was determined by comparing the mean total score made by persons who answered the item correctly with the mean total score made by those who did not. The efficiency of the alternative (wrong) answers as "distractors" was also assessed from these data.

A secondary criterion of a satisfactory item was the proportion of those attempting it who answered correctly. For each subtest, a distribution of item-difficulty was sought in which there would be a few easy and a few difficult items and a concentration around the midpoint between chance and perfect proportions.

In general, a satisfactory item was considered to be one which showed:

a. a reasonable number of omissions;
b. a mean subtest score for those persons selecting the correct response which was higher than the general mean for the subtest;
c. a set of efficient distracting alternative answers.

Form 1 was intensively studied after being put into general use at all training centers. The knowledge gained in these studies was helpful in the preparation of Forms 2 and 3, which were analyzed in preliminary form. On Forms 2 and 3 of the mechanical tests, and on the Fleet Edition of the Basic Test Battery (with the exception of the Clerical Aptitude Test), item analyses were made by comparing the upper and lower 27 per cent of the distribution on each test.[6]

TIME LIMITS. For the determination of optimal time limits for each test, experimental studies were carried out with the preliminary

[6] See Flanagan, J. C. "General Considerations in the Selection of Test Items," *Jour. Educ. Psychol.*, 1939, 30, pp. 674-680.

forms. Each of these forms was given to a carefully selected sample of recruits at a naval training station. During the administration of each test the recruits were required to indicate on each test or subtest the progress made at the end of a chosen period of time. Three time limits were chosen: a long time period which should allow nearly all the recruits to finish, and two shorter periods. The procedure was for the experimenter to call "Mark" at the end of the short period, and have the recruits mark their papers at the item on which they were then working. After an extension of time the experimenter called "Mark" again and the recruits again indicated their progress. The third limit was arrived at by another extension of time.

Scores on the first and second time limits were correlated with the maximal score obtained by the greatest time allowance. The proportion of recruits completing each test or subtest within each time limit was also determined. The time limits to be allowed for each test in final form were determined on the basis of the following criteria:

a. the time limit for which scores correlated highly with those obtained under the maximal time allowance;
b. a time limit under which approximately 50 per cent of the subjects finished the test;
c. a time limit for which there was a well-balanced distribution of scores.

Normative Studies

NORMS. In the establishment of norms for the Basic Test Battery, a standard scale was devised on which the mean score for each test was assigned a value of 50 and the standard deviation a value of 10. Scores from all training centers were reported in terms of this Navy Standard Score (NSS). The groups from whose data the norms were established were chosen to be representative of the entire recruit population. All training centers were included, each in proportion to its contribution to the national intake of recruits. Norms for all forms of each of the tests have been established in a similar manner.

Norms for the Fleet Edition of the tests were derived in a similar manner by administering the tests of this edition at a naval training station together with Forms 1 and 2 of the Basic Test Battery. The technique employed with the General Classification Test was as follows:

Two groups of recruits were chosen. One group took the experimental test of the Fleet Edition and Form 2 of the Basic Test Battery. The other group took Forms 1 and 2 of the Basic Test Battery. From the results of these administrations it was possible to compare Form

1 with the Fleet Edition test which was intended to be parallel to it. These two forms could not be given to a single group, inasmuch as the contents of the two tests were in large part identical. Statistical procedures were then used to equate the scores of Form 1, Basic Test Battery with the scores of the Fleet Edition test.

COMPARISON OF STATIONS. It had been noticed that differences existed in scores made by recruits at different naval training stations on the tests which were in use prior to the introduction of the Basic Test Battery. A study of these tests, made in early 1943, had shown that some of the stations regularly received recruits of apparently higher caliber than those in other stations. Since each station regularly reported the mean, range, and standard deviation for its recruits on all tests of the Basic Test Battery, it was possible to compare

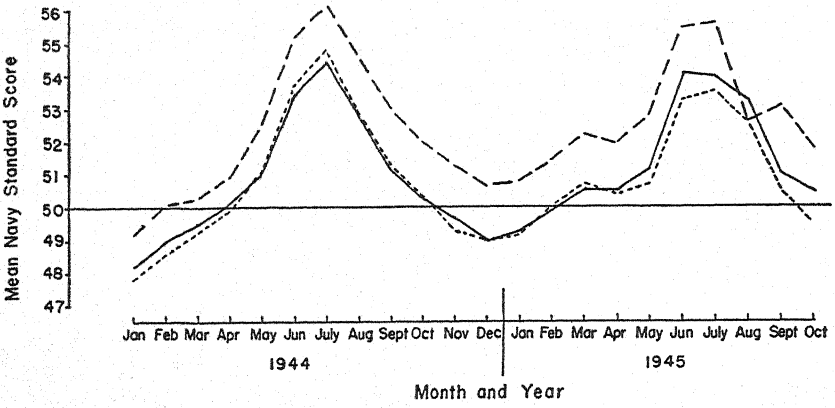


Figure 5-VI. Monthly trends in scores in verbal and mathematical tests of the Basic Test Battery.

the recruits in different stations, and therefore, in different parts of the country over a long period of time. The results of the comparison of five stations are shown in Table 1-VI.

TIME TRENDS. Since each station reported its scores bi-weekly, it was possible to observe the trends in mean scores over a long period of time. The results of the study of these trends are shown in Figures 5-VI and 6-VI. It can be seen in these figures that, although there are some differences among the tests, and there is a difference between the two years shown in the figures, the general trend is for high mean scores in June and July and low scores toward the end of the year.

Several factors undoubtedly contribute to this trend, but prominent among them are probably the following: June and July are the months in which high school graduates entered the services.

TABLE 1-VI. Mean scores on Forms 2 and 3 of the Basic Test Battery for five Naval Training Centers

Test	Station										
	A [1]		B [2]	C [3]		D [4]		E [5]		All Stations [6]	
	Form 2	Form 3	Form 2	Form 2	Form 3	Form 2	Form 3	Form 2	Form 3	Form 2	Form 3
General Classification Test	49.1	48.4	51.5	51.2	53.5	50.6	51.2	48.7	51.3	50.5	51.8
Reading Test	50.4	50.1	53.2	52.7	55.6	51.7	51.9	49.1	53.1	51.8	53.4
Arithmetical Reasoning Test	49.2	48.6	51.5	51.7	54.2	49.6	49.8	48.4	49.8	50.4	51.4
Mechanical Aptitude Test	50.4	50.9	52.7	52.0	54.0	49.5	50.9	49.5	53.2	51.0	52.7
Mechanical Knowledge Test (Mechanical Score)	49.6	47.1	53.5	51.6	50.8	48.3	47.6	49.7	50.6	50.6	49.5
Mechanical Knowledge Test (Electrical Score)	49.7	48.4	52.0	51.8	54.1	48.4	49.7	48.8	51.2	50.4	51.6

[1] For Form 2, N = 84,107; for Form 3, N = 37,576.
[2] For Form 2, N = 108,701 to 108,875.
[3] For Form 2, N = 286,860 to 288,034; for Form 3, N = 120,005 to 120,865.
[4] For Form 2, N = 178,327 to 178,328; for Form 3, N = 72,835.
[5] For Form 2, N = 116,035 to 117,730; for Form 3, N = 70,956 to 71,333.
[6] For Form 2, N = 776,884 to 774,705; for Form 3, N = 302,526 to 301,372.

During these months, also, the men who failed as aviation cadets in the V-5 program, but who were above the average on the aptitude tests, reported at the naval training centers and took the Basic Test Battery.

It is also evident from Figure 5-VI that the scores for 1945 on some of the tests, notably the Mechanical Knowledge Test (Mechanical Score) were lower than in the preceding year. This was presumably due to the fact that recruits in the latter year were in general younger than in 1944. It has repeatedly been found that age is positively correlated with scores on mechanical information tests. That

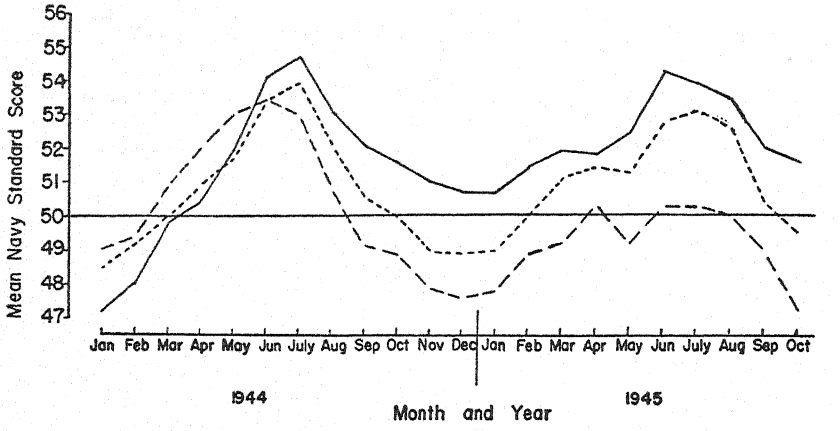


Figure 6-VI. Monthly trends in scores on mechanical tests of the Basic Test Battery.

correlation probably accounts for the trend of this mechanical score over the two years.

Test Reliabilities

The reliabilities of Forms 1, 2, and 3 of the general tests of the Basic Test Battery have been determined by the split-half method using representative samples of answer sheets routinely submitted to the Test and Research Section. For each test, total scores on the odd-numbered and on the even-numbered items were computed separately and correlated. The resulting coefficients of correlation were corrected by the Spearman-Brown Prophecy Formula to obtain estimates of the reliabilities of the full-length tests. The Kuder-

Richardson Formula No. 21 has also been used in calculating reliability coefficients for these tests and for the Clerical Aptitude and Spelling tests. The estimated coefficients of reliability are shown in Table 2-VI.

Inspection of Table 2-VI reveals that three of the tests, the General Classification Test, the Mechanical Aptitude Test, and the Mechanical Score of the Mechanical Knowledge Test are highly reliable, with odd-even coefficients of .90 or more on all forms.

The Arithmetical Reasoning Test has a satisfactory reliability on all forms, especially in view of the small number of items (30) in the

TABLE 2-VI. Estimated reliability coefficients of Basic Test Battery based upon data obtained from routine administration to recruits [1]

Test	Reliability of Each Form [2]			Alternate-form Reliabilities [3]	
	Form 1	Form 2	Form 3	Forms 1 & 2	Forms 2 & 3
General Classification Test	.91(200)	.95(200)	.96(250)	.90(450)	.93(500)
Reading Test	.82(200)	.89(200)	.87(250)	.72(450)	.81(500)
Arithmetical Reasoning Test	.86(200)	.88(200)	.90(250)	.82(450)	.86(500)
Mechanical Aptitude Test	.95(200)	.95(200)	.95(200)	.88(250)	.87(500)
Mechanical Knowledge Test (Mechanical Score)	.90(200)	.92(200)	.91(200)	.87(250)	.86(500)
Mechanical Knowledge Test (Electrical Score)	.84(200)	.89(200)	.82(200)	.78(250)	.83(500)
Clerical Aptitude Test	.91(780) [4]	.95(400) [4]		.86(400)	
Spelling Test	.85(780) [4]	.91(400) [4]		.73(400)	
Radio Code Test	.88(470) [5]				

[1] Figures in parentheses represent the number of cases in the samples used for computing reliabilities.

[2] Spearman-Brown Formula used to estimate reliability from correlation between odd and even-numbered items except where indicated by footnotes 4 and 5.

[3] Based on the correlations between alternate forms of the same tests.

[4] Reliability coefficients computed by Kuder-Richardson Formula No. 21.

[5] Spearman-Brown Formula used to estimate reliability from correlation between Part I and Part II.

test. A relatively small increase in the length of this test would suffice to raise its reliability above .90. Such a change was introduced in the new experimental form, X-4, by the addition of a test of arithmetical computation to the arithmetical reasoning items. Data regarding this new Arithmetic Test are not yet available, but it is confidently expected that the reliability will be higher than on the first three forms.

The reliability coefficients of the Reading Test and the Electrical Score of the Mechanical Knowledge Test are satisfactory for screening purposes, but should be higher if the tests are to be used for

assigning personnel to one school rather than another. This criticism applies particularly to the Mechanical Knowledge Test.

The alternate-form reliabilities of all tests are lower than the odd-even estimates. This result is to be expected in view of the fact that a test typically correlates with itself higher than with another test. In the case of the correlation between Forms 1 and 2 of the Reading Test, Spelling Test, and Mechanical Knowledge Test (Electrical Score) the correlations, all below .80, indicate that the two forms were not very exactly matched.

BASIC TEST BATTERY (FLEET EDITION). The tests of the Fleet Edition of the Basic Test Battery were intended to parallel the forms used in training centers. Estimates of the reliabilities of the tests in the Fleet Edition were computed by the Kuder-Richardson Formula

TABLE 3-VI. Reliability coefficients of Basic Test Battery (Fleet Edition) based on data from recruit performance

Test	Estimated Reliability Coefficients; Basic Test Battery (Fleet Edition) Form 1[1]	Correlations of Tests of Basic Test Battery (Fleet Edition) with Comparable Tests of Basic Test Battery, Form 1
General Classification Test	.90(500)[2]	.84(500)
Arithmetical Reasoning Test	.77(500)	.86(500)
Mechanical Aptitude Test	.88(500)	.74(500)
Mechanical Knowledge Test (Mechanical Score)	.87(500)	.64(500)
Mechanical Knowledge Test (Electrical Score)	.79(500)	.66(500)
Clerical Aptitude Test	.96(400)	.78(400)

[1] Reliability coefficients computed by Kuder-Richardson Formula No. 21.
[2] Figures in parentheses represent number of cases in sample used.

from an administration to recruits who also took the tests of Form 1 of the regular battery.

In Table 3-VI are shown the estimated reliability coefficients of the Fleet Edition tests and the correlation coefficients of these with the comparable tests of Form 1. The reliability coefficients for the tests of the Fleet Edition are all lower than those for Form 1 of the Basic Test Battery. Four of the six coefficients for the Fleet Edition are satisfactory or better, while the other two are not.[7] The lower reliability of the Electrical Knowledge Test can be accounted for completely on the basis of the smaller number of items, but the Arithmetical Reasoning Tests of both forms contain the same num-

[7] It has been customary in the work on the Navy's aptitude testing program to consider that for tests of the type here described, reliability coefficients below .80 are not acceptable, those from .80 to .89 are satisfactory, and those of .90 or above are very high.

ber of items (30). It is therefore likely that the fleet form of the Arithmetic Test contained a number of items that are intrinsically unreliable.

The extent to which the tests of the Fleet Edition match those of Form 1 is also shown in Table 3-VI. Due again to the disparity in the number of items, and to the differences in time limits, the correlation coefficients are not as high as should be expected of two tests on which a large number of items are identical.

Relation of Test Scores to Age and Educational Level

Coefficients of correlation between test scores and age and between test scores and highest school grade completed were computed for

TABLE 4-VI. Correlation coefficients of scores on Basic Test Battery with age and with highest school grade completed

Test	Correlation Coefficients: Test Score with Age (N = 906)	Correlation Coefficients: Test Score with Highest School Grade Completed (N = 908)	Navy Standard Score: M	Navy Standard Score: σ
General Classification Test	.10	.65	50.1	10.6
Reading Test	.02	.58	51.0	11.7
Arithmetical Reasoning Test	.12	.57	50.0	11.6
Mechanical Aptitude Test	—.02	.52	50.4	10.3
Mechanical Knowledge Test (Mechanical Score)	.30	.39	52.1	11.2
Mechanical Knowledge Test (Electrical Score)	.19	.56	50.8	10.4
Clerical Aptitude Test	—.09	.65	50.3	9.2
Spelling Test	.06	.53	50.1	9.9
Radio Code Test—Speed of Response	.05	.41	52.3	10.9
Age		—.02	26.1	6.7
Highest School Grade Completed	—.02		10.2	2.3

a national sample of recruits in April 1944. The correlations are shown in Table 4-VI. It is clear from this table that the correlation coefficients between test scores and educational level are moderately high, as is to be expected. The correlation coefficients between test scores and age are, with two exceptions, close to zero. The exceptions are the two scores on the Mechanical Knowledge Test, which, as previously stated, has been found repeatedly to be correlated positively with age. It is noteworthy that of all test scores of the Basic Test Battery, only these two, from an information rather than an aptitude test, are correlated with age. The inference is that as one

grows older he may acquire a background of information relevant to mechanical and electrical tools and processes, but does not improve in any of the verbal abilities tested, in arithmetical processes, or in the special aptitudes which are included in the battery.

Intercorrelation Studies

Two kinds of study are presented as the result of intercorrelating the tests of the Basic Test Battery in Forms 1, 2, and 3. Tables of intercorrelations were prepared in which the relationships of the tests could be observed, and some of these tables were subjected to factor analyses in order to disclose the relationships among the tests, and what factors were represented by the tests.

INTERCORRELATIONS. The intercorrelations among the tests of the Basic Test Battery, Forms 1, 2, and 3, are shown in Table 5-VI. The data for Form 1 of the tests were obtained from a representative sample of 500 recruits from all naval training stations from July to October 1943. Data for Form 2 were collected from a sample of recruits in all naval training stations in April 1944. The data for Form 3 were obtained from a similar sample in January 1945. On all three samples the means and standard deviations were very close to those of the national recruit population at the time.

It can be seen from Table 5-VI that all of the tests are positively related. The highest correlation coefficients were between the General Classification and Reading Tests (.80 to .85), which were expected in view of the fact that both are tests of verbal ability. The lowest correlation coefficients were found between the Mechanical Score of the Mechanical Knowledge Test, Form 2, and the scores on Form 1 of the Clerical Aptitude, Spelling and Radio Code Tests (.36, .33, .34).

The Arithmetical Reasoning Test and the three scores on the mechanical tests correlate higher with the verbal tests than is desirable. The Mechanical Score of the Mechanical Knowledge Test appears to be the one mechanical test score which is most independent of verbal ability. The correlation coefficients between this score and the verbal test scores range from .39 to .57.

The two scores obtained from the Mechanical Knowledge Test are quite highly correlated (.67 to .78). An analysis of the tests showed that the form in which the items were cast was an important factor in this correlation. The items of both parts of this composite test which are alike in form, although different in content, correlated more highly than the two component types of item in each score. The mechanical and electrical scores are each derived from both verbal and pictorial items. The correlation coefficients of mechanical-

pictorial items with electrical-pictorial, and of mechanical-verbal with electrical-verbal, were as high as, or higher than those of mechanical-pictorial with mechanical-verbal items or of electrical-pictorial with electrical-verbal items. It was partly because of the intercorrelation between the two scores of the Mechanical Knowledge

TABLE 5-VI. Intercorrelations among tests of the Basic Test Battery based on data from routine administration to recruits [1]

	Test	Variable B	C	D	E	F	G	H	I	Navy Standard Score M	σ
A	General Classification										
	Form 1	.81	.69	.60	.49	.53				47.50	10.21
	Form 2	.85	.79	.69	.57	.73	.68	.68	.61	50.13	10.53
	Form 3	.81	.72	.60	.43	.55				52.20	10.41
B	Reading										
	Form 1		.69	.56	.46	.51				45.00	10.00
	Form 2		.77	.67	.53	.68	.64	.63	.51	51.15	11.61
	Form 3		.68	.61	.50	.60				54.09	10.85
C	Arithmetical Reasoning										
	Form 1			.61	.41	.47				45.00	9.95
	Form 2			.69	.53	.68	.64	.58	.51	49.99	11.62
	Form 3			.59	.41	.50				51.04	11.16
D	Mechanical Aptitude										
	Form 1				.55	.53				48.00	10.31
	Form 2				.61	.69	.65	.48	.49	50.51	10.28
	Form 3				.59	.60				52.95	9.82
E	Mechanical Knowledge (Mechanical Score)										
	Form 1					.78				48.00	10.55
	Form 2					.75	.36	.33	.34	52.32	11.19
	Form 3					.68				49.67	9.77
F	Mechanical Knowledge (Electrical Score)										
	Form 1									48.00	9.47
	Form 2						.55	.54	.45	50.89	10.43
	Form 3									51.56	9.62
G	Clerical Aptitude										
	Form 1							.66	.53	50.24	9.08
H	Spelling										
	Form 1								.39	50.12	9.89
I	Radio Code										
	Form 1									52.38	10.95

[1] The data for Form 1 are based on a sample of 500 persons for whom scores were obtained on Form 1 of the General Classification Test, Reading Test, Arithmetical Reasoning Test, Mechanical Aptitude Test, Mechanical Knowledge Test (Mechanical Score), and Mechanical Knowledge Test (Electrical Score). The data for Form 2 were obtained from a sample of 933 persons for whom scores were obtained on Form 2 of the tests listed above plus Form 1 of the Clerical Aptitude Test, Spelling Test, and Radio Code Test. The data for Form 3 are based on a sample of 803 persons for whom scores were obtained on Form 3 of the general tests of the battery.

Test that an extensive revision of the format of this test was undertaken for Forms 4 and 5. Since the Electrical Score was quite highly correlated with the Mechanical Score, the advantage of using both scores was questionable. Since the reliabilities of one of the scores (Electrical) was not particularly high (.82 to .89), and the reliability of the difference between the two scores was low, the classification of individual men based on this difference was subject to rather large errors. Therefore, in the new experimental forms of the Mechanical Test an effort was made to increase the reliability of the test, and the verbal items were deleted. In their place, the Mechanical Comprehension items from the Mechanical Aptitude Test were substituted. Data are not available at present on the reliability of the new forms.

FACTOR ANALYSES. Two factor analyses have been performed on the tests of the Basic Test Battery, the first on Form 1 of the general tests, and the second on Form 2 of the general tests plus Form 1 of the Clerical Aptitude, Spelling, and Radio Code Tests. The first analysis revealed three factors, one clearly verbal and best represented by the General Classification, Reading, and Arithmetical Reasoning Tests, the second mechanical and best represented by the Mechanical Knowledge Test, and a third factor which was not well defined in the analysis, but on which two parts of the Mechanical Aptitude Test had relatively high loadings. The three factors appeared relatively independent, the highest correlation (between the mechanical and the third factor) being .31 and the lowest correlation (between the verbal and mechanical factors) being .19. It is noteworthy that this analysis failed to disclose any independent quantitative reasoning factor represented by the Arithmetical Reasoning Test.

The second analysis, including all of the special tests of the battery, revealed four factors. The first was again a general (verbal) intelligence factor, the second a mechanical information factor, and the third and fourth were not well defined.

The loadings of all tests on the first factor (A) are .57 or higher, and the verbal tests and Arithmetical Reasoning Test have high loadings (.86 or higher) on this factor and very low loadings (.11 or less) on the other three. From this it appears that Factor A is a generalized factor of verbal intelligence, which is correlated with all the other factors.

Both the Mechanical and Electrical Scores of the Mechanical Knowledge Test have high loadings (.50 or more) on the second factor (B) and zero loadings (.11 or less) on Factors C and D. From the fact that this factor is well represented only on the Mechanical Knowledge Test, it is defined as a specialized mechanical information factor.

Factor C was represented, as in the first analysis, primarily on the Mechanical Aptitude Test, and in this analysis on the Radio Code Test. Since the Block Counting and Surface Development parts of the Mechanical Aptitude Test involve the ability to perceive visual relationships, and the Radio Code Test involves the ability to perceive auditory relationships, Factor C has been called tentatively "Speed in perception of visual and auditory relationships."

Factor D is also not well defined in the analysis. The Spelling Test is the only one which has moderate loadings on it and zero loadings on Factors B and C. The Clerical Aptitude Test has loadings of .30 on both Factors C and D.

Validity of the Basic Test Battery[8]

Since Form 1 of the Basic Test Battery was put into national use in June 1943, a continuing program of studies of the validity of the tests has been carried on to insure that the tests are used as effectively as possible. In these studies the validating criterion has been success in service schools.

The usual procedure employed in the study of the validity of the Basic Test Battery in predicting success in service schools has been to compute correlation coefficients between the trainees' service school grades and the scores on one, and on combinations of two or more tests of the Basic Test Battery. In the latter case multiple correlation coefficients and correlation coefficients based on the summed scores of the two (or more) tests have been used.

It has been found in these studies that every one of the six scores obtained from administration of the general tests of the Basic Test Battery has been useful in classifying men to some school. The best single test, and the best combination of two tests, for predicting success in different schools, has, of course, varied from school to school. The degree of relationship has also varied, but in general the most valid single test score has correlated approximately from .40 to .65 with school grades, and the best combination of two tests approximately from .45 to .70 in terms of multiple correlation.[9] The

[8] For a more detailed discussion of the validity of the Basic Test Battery, see Chapter XII.

[9] In computing the validity coefficients of these tests, corrections were first made for the restricted dispersions of scores in the samples used. The coefficients were based on the dispersion of scores in the total recruit population for which the standard deviation was known to be 10. The formula for this correction for curtailment has been given by Kelley, Truman L., *Statistical Method,* Macmillan, New York, 1923. The second statistical procedure consisted of averaging the various coefficients obtained at a sample of schools by means of the z transformation of Fisher. The procedure followed in transforming to z and converting the average z to r has been described by Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases,* McGraw-Hill, New York, 1940.

correlation coefficient between school grades and the summed scores on two tests were generally .01 to .03 lower than the multiple correlation coefficients.

Extensions of the Basic Test Battery

Aside from the Fleet Edition of the Basic Test Battery, which has already been described, two additional tests were developed for use with enlisted personnel, one to be used in addition to the basic battery, and the other as a substitute for one of the tests in the battery. These extensions of the Basic Test Battery were the Applicant Qualification Test and the Arithmetical Computation Test.

APPLICANT QUALIFICATION TEST. The Applicant Qualification Test was developed for use in the intellectual screening of 17-year old applicants for naval service. The test was designed (1) to require less than one hour of testing time, (2) to be suitable for hand-scoring, (3) to be printed in expendable booklets, (4) to yield a single score which would be convertible to Navy Standard Scores for purposes of comparison, (5) to be capable of administration, scoring, and interpretation by relatively untrained personnel, and (6) to be available in several forms.

It was decided that in order to meet the needs of the Navy for various abilities the Applicant Qualification Test should be a composite test, measuring arithmetical and mechanical abilities as well as verbal. Since the test was to be used for screening, and would be used to eliminate only those at the bottom of the distribution of scores, it was decided to use a preponderance of easy items in order to obtain a negatively skewed distribution of scores. Thus the finest discrimination would be obtained where it was most needed, at the lower end of the scale.

Description. The Basic Test Battery functioned as the primary source of supply of items for the Applicant Qualification Test. The verbal section consists of analogies and sentence completion items, the mechanical section consists of mechanical and electrical information items, and the remainder of the test consists of arithmetical reasoning items. The use of these three types of item permitted the test to be uniform in format and quite simple, a desideratum in view of the fact that the test would probably be the first of its kind encountered by many of the applicants for Navy service. On page 77 is a tabular description of the Applicant Qualification Test, for which the time limit was forty-five minutes.

Administration. An experimental form containing 100 items was administered to 412 recruits at a naval training center. On the basis of an item analysis, the 80 best items were retained for Form 1. This

Test	Number of Items
Applicant Qualification Test, Form 1	
Analogies	16
Sentence Completion	16
Arithmetical Reasoning	20
Mechanical Information	16
Electrical Information	12
Total	80

test was then given to a sample of 500 recruits, with whom it was found that the 45-minute time limit permitted the majority to complete the test.

Form 1 of the Applicant Qualification Test was then administered to 509 recruits who comprised the standardization group. One-half took the test shortly before the Basic Test Battery, and the other half shortly after taking the Basic Test Battery.

The mean raw score in the test was 49.13 and the standard deviation was 16.28. The distribution of scores showed a slight negative skewness, which indicates good discrimination on the test in the region where it is desired.

Reliability. The reliability of the test, calculated by the Kuder-Richardson Formula No. 21, was found to be .90, which is a slight underestimation owing to the assumptions of complete homogeneity and equal difficulty of items which underlie the Kuder-Richardson Formula. An odd-even reliability coefficient would probably be somewhat higher than .90.

Validity. Estimates of the probable validity of the Applicant Qualification Test for screening the 17-year old applicants for naval service were made by noting the correlation coefficients between it and the Basic Test Battery. It was found that the highest correlation coefficient was between the Applicant Qualification Test and the General Classification Test (.85), while the others ranged from .57 to .80. It is felt that insofar as the Basic Test Battery scores are valid predictors of performance in naval training schools, the Applicant Qualification Test shows promise of effectiveness, owing to its correlations with the Basic Test Battery.

ARITHMETICAL COMPUTATION TEST. As a result of finding that the Arithmetical Reasoning Test correlated quite highly with the verbal tests in the basic battery, an attempt was made to devise an arithmetic test relatively free of the verbal factor. To this end the Arithmetical Computation Test was constructed. In this test the attempt was made to have the scores reflect more the recruits' ability to manipu-

late numbers and less the ability to read directions and manipulate words. Accordingly, the directions were made as simple as possible, the testees being instructed merely to add, subtract, multiply or divide the numbers given them.

Construction and Analysis. An experimental form of the test was constructed of sixty items of free-answer form. In the test the four fundamental operations of arithmetic were about equally represented. The test called for the ability to manipulate whole numbers, decimals, and fractions. This test was administered to 1,430 recruits at a naval training center, and an item analysis was performed on a random sample of the answer sheets. The wrong answers most commonly given to the items selected for retention provided the distractors for the multiple-choice form of the test which was issued later. This edition of the test was given to another sample of 1,000 recruits and was item-analyzed.

Statistical Characteristics. A study of the statistical characteristics of the first multiple-choice form of the test, Form X-2, revealed that it had a satisfactory difficulty and reliability. The fifty-item test had a mean of 24.2, a standard deviation of 11.9, and a reliability coefficient of .95. The items were uniformly of high discriminative value, the median biserial coefficient of correlation being .70.

Correlational analysis showed, however, that the scores on the Arithmetical Computation Test correlated as highly with General Classification Test scores (.76) as did the scores on the earlier test of arithmetical reasoning (.69 to .79). From this fact it was concluded that the Arithmetical Computation Test failed to fulfill an important objective, namely, a low correlation with scores on the verbal tests of the Basic Test Battery. Despite this fact, the Arithmetical Computation Test was considered to be of promising utility because of its high reliability and the fact that it appeared to be about as valid as the tests of the Basic Test Battery for predicting success in a number of schools.

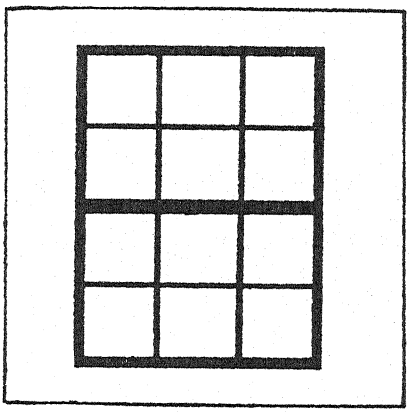
Selection Tests Developed for the Literacy Training Program

When the Navy established a training program for non-readers and poor readers, it became necessary to develop tests which (1) would ascertain the reading level of the poor readers, and (2) identify those individuals who, while they were poor in reading skills, possessed enough general mental ability to profit from a special training program in reading. The Literacy Test and the Non-Verbal Classification Test were developed for these purposes.

LITERACY TEST. The Literacy Test Form X-1 was developed to measure reading achievement. It consists of 44 word recognition and

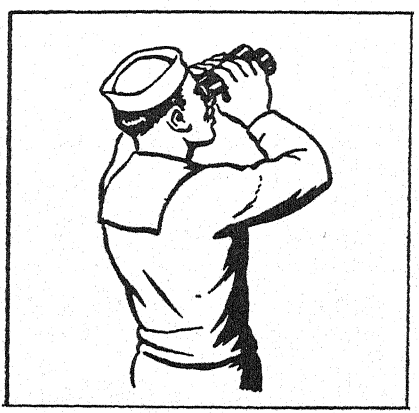
reading comprehension items of the type illustrated in Figure 7-VI. It will be noted that the items are constructed from materials drawn from the experience of naval personnel. The items are arranged roughly in order of difficulty, and no time limit is used. The exam-

Sample A. A word-recognition item. The test contains 20 of these items.



light
window
hinge
blind
windy

Sample B. A sentence-reading item. The test contains 16 of these items.



There is a ship in sight.
The sailor is looking upwards at something.
The glasses are very powerful.
A battle is about to begin.

Sample C. Paragraph reading. The test contains 8 of these items.

Each recruit is given a complete outfit of clothing when he enters the Navy. He is also given a stencil for marking his clothes. If he takes proper care of his outfit, it will last for a long time; and if he marks everything with his name, he won't lose anything.

A recruit
should buy his clothes before he enters the Navy.
should make the clothes given to him last until the end of the war.
will lose a month's pay if he doesn't mark his clothes.
should take proper care of the clothes that are given to him.

Figure 7-VI. Literacy test sample items.

inee records responses by encircling the word, phrase, or statement which he believes to be correct.

The Literacy Test was standardized by administering it and the Gates Reading Survey Test to two companies of recruits at a naval training center and to two companies in the special reading training program. On the basis of data thus obtained, a conversion table was constructed for converting raw scores in the Literacy Test into reading grade levels. Recruits scoring below fifth-grade reading level and possessing sufficient mental ability to profit from special instruction were to be transferred to the special training program in reading.

At the time of the termination of hostilities, Form X-2 had been developed and was in the process of standardization. This form contains 60 items distributed as follows: word recognition, 20; sentences, 15; paragraph reading, 25. The directions for Form X-2 are given orally instead of printing them in the test booklets. The response technique was changed from encircling responses to checking boxes placed opposite the choices.

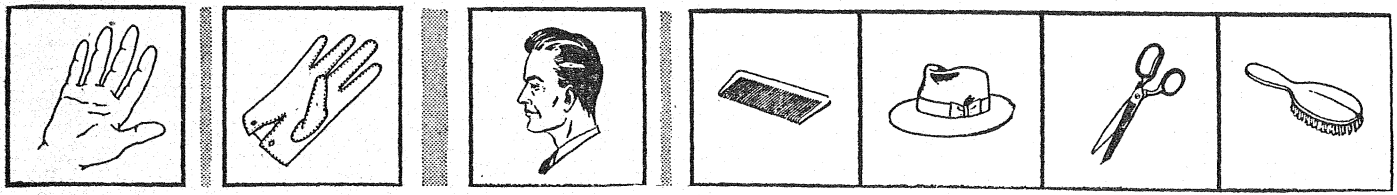
NON-VERBAL CLASSIFICATION TEST. This test was developed to measure the learning ability of recruits who scored low in the Reading or General Classification Tests of the Basic Test Battery, and to determine whether they possessed sufficient mental capacity to profit from instruction in reading. In building the test, two important facts were kept in mind: (1) the test would be used with adults, and (2) it would have to be given to large groups of men. Furthermore, it was desired that a man's score in the test should serve in the place of the General Classification Test score as the index of his general mental ability.

Three types of items were chosen for use in the test: analogies, identification of the "non-belonging" figures in a series, and completion. Each of these is illustrated in Figure 8-VI. The examinees recorded their responses by crossing out the choice they believed to be correct.

The test was administered in experimental form to both literate and illiterate groups at a naval training center. The first group consisted of 200 representative literates who also took Form 2 of the General Classification Test. The coefficient of correlation between distributions of raw scores in the two tests was .74. The second group consisted of 1,071 illiterates. The means and standard deviations in the test (100 items) for the two groups were as follows: literate 56.8 and 16.5; illiterate 46.2 and 11.4. After the item analysis was completed the number of items in the test was reduced from 100 to 75 and a conversion table prepared for converting raw scores to equivalent scores in the General Classification Test, Form 2.

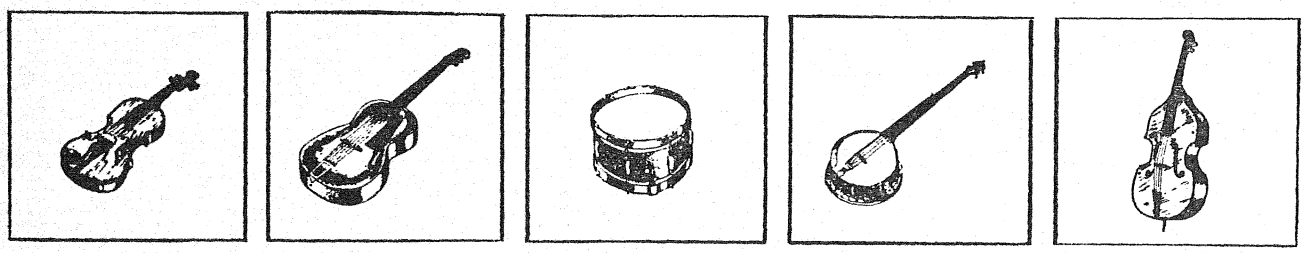
Sample A. Analogies type item.

The test contains 30 of these items.



Sample B. Non-belonging type item.

The test contains 20 of these items.



Sample C. Completion type item.

The test contains 25 of these items.

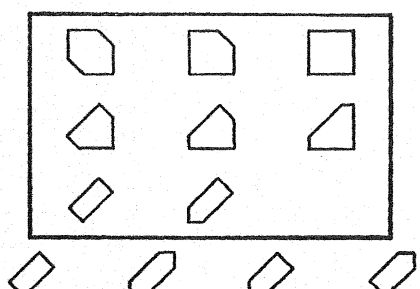


Figure 8-VI. Non-verbal classification test sample items.

Revisions and Future Developments

In the course of the studies of Forms 1, 2, and 3 of the Basic Test Battery, several instances of failure of the tests to perform optimally became evident. For example, from the intercorrelation studies, including the factor analyses, it was evident that the Arithmetical Reasoning Test was highly related to the verbal tests and that no independent quantitative or numerical reasoning factor was represented by any of the tests. The Arithmetical Computation Test represented an attempt to reduce to a minimum the required verbal intelligence, but it correlated as highly with the verbal measures as did the Arithmetical Reasoning Test. The Arithmetic Test in Forms 4 and 5 consists of both computational and arithmetical reasoning items.

Intercorrelation studies also showed a high degree of relationship between the General Classification Test and the Reading Test. Since the General Classification Test had a higher reliability and seemed somewhat superior to the Reading Test in the prediction of success in service schools, it was decided to retain the General Classification Test as the measure of verbal ability in Forms 4 an 5 and to omit the Reading Test.

In the measures of mechanical abilities in Forms 1, 2, and 3, two scores were derived from the Mechanical Knowledge Test (Mechanical Score and Electrical Score) in order to make differential classifications to electrical or mechanical schools. For the purpose of differential classification both tests would need high reliability in order that the differences between scores be reliable. The reliability of the electrical score was not high enough to permit dependable classification to be based on the difference between the two scores of mechanical knowledge.

In the experimental editions of Forms 4 and 5 of the Mechanical Test of the Basic Test Battery the attempt has been made to increase the reliability of the test and to decrease its correlation with the other tests of the battery. The new Mechanical Test consists of mechanical and electrical knowledge items, pictorially presented, and mechanical comprehension items. Instead of the three mechanical test scores provided by Forms 1, 2, and 3, there will be but one mechanical score. At the present time no experimental evidence is available as to the effectiveness of this test.

Of the six general tests in Forms 1, 2, and 3, the three which are being retained in Forms 4 and 5 are the General Classification Test, Arithmetic Test, and Mechanical Test. Of the three special tests, only the Clerical Aptitude Test is being retained. The Spelling Test

is being deleted entirely and the Radio Code Test will be administered as a special aptitude test.

In the interest of processing enlisted men at the highest possible speed during the war, all men were given the entire battery of basic tests. At that time it was felt, however, that other procedures, more economical of testing time, might be preferable for long-term purposes. Among such possible programs, consideration should be given to the possibility of a multiple-stage testing program which would utilize a short primary battery, given to all men for screening purposes, and additional secondary batteries to be given to smaller numbers of men for the purpose of classification for specific schools. Related problems concern the smallest number of tests which can be used effectively in the primary battery, and what special tests are needed in the secondary stage.

Thought is also being given to the possible use of other types of tests in the secondary stage of testing. Among the types being considered are proficiency tests, tests of interests and personal adjustment, and aptitude tests of the performance type. A more extended discussion of these and related problems will be found in Chapter XXII.

CHAPTER VII

BASIC TESTS FOR OFFICER PERSONNEL

OWING to the very rapid expansion of the Navy in 1942, there was need for a general purpose test which could be used in the selection of officer candidates. At that time large numbers of officers were taken into the Navy directly from civilian life through the offices of naval officer procurement. They were then sent to indoctrination schools for training. The first test for officer candidates was constructed for use at the offices of naval officer procurement for the purpose of screening out those persons who probably would not succeed in Navy training programs or billet assignments. Subsequent tests were devised to be administered at primary officer schools or classification centers to provide differential information regarding more specific aptitudes of officer candidates. This differentiating information could be used in classification for further training or for specific ship or shore billets.

The principal criterion against which the success of any selection program in the Navy should be measured is, of course, an individual officer's success in the performance of his duties. The final validity of the tests developed for selection and classification should, therefore, be estimated only after correlating the test results for a large number of individuals with accurate and reliable appraisals of their performance on the job.

The lack of time available for the development of workable selection tests for naval officer candidates, however, precluded any validation of the tests by performance criteria. But, since most officer candidates accepted at offices of naval officer procurement were sent to some school for training, the selection tests to be used at these offices could be validated against achievement in school and could be used at once in an effort to reduce the proportion of training failures. The first aim of a selection test was, therefore, to screen out the candidates who would probably not succeed in Navy training programs. For this purpose the Officer Qualification Test (OQT) was developed. The test developed for administration during the period of primary training, for the purpose of classifying officers for advanced training for specific billets was the Officer Classification Test (OCT). This test will be discussed in the second part of this chapter. The third part contains a brief discussion of the College Qualifying Tests devised to select candidates for the V-12 college training program.

Since the Officer Qualification Test was the first of the basic tests developed for officers and since it served as the model for the development of later tests, it has been more extensively studied than any other, and hence is described in greater detail in this report.

Officer Qualification Test

In view of the fact that the Officer Qualification Test was to be given at many different places by personnel not uniformly trained for the purpose, one of the first specifications set up for the test was that it be virtually self-administering. The use of a self-administering test would obviate the need for trained testing personnel or for rigid standardization of the procedures to be employed in giving the test.

The limited testing time available at the offices of naval officer procurement imposed the additional requirement that the Officer Qualification Test take not more than one hour. It was specified, however, that it be a power test. This meant that even if the time-limit were removed, the variability of test scores should remain large. Particularly, it meant that with the time-limit of one hour the scores should not reflect merely speed in making decisions and recording answers.

Until late in 1942, no single general aptitude test had been uniformly used at the various offices of naval officer procurement in the selection of candidates for commissions. By December of that year, an experimental form of an officer qualification test, called the Officers' Selective Examination (SE O-1) was developed and put into use at the New York Office of Naval Officer Procurement. On the basis of an analysis of a sample of answer sheets from that office, the Officer Qualification Test, Form 1, was prepared and used nationally.

After Form 1 had been put into use at all offices of naval officer procurement, the preparation of two parallel forms of the Officer Qualification Test (Forms 2 and 3) was begun. The first step in the preparation of Forms 2 and 3 was to devise a lengthy experimental test (Form O-2) by adding to Form 1 enough items to provide for two parallel forms, with a number of excess items which could be discarded. Form O-2 was administered to a group of male officer candidates at a naval indoctrination school and to a group of women at a reserve midshipmen's school for WAVES officers. From the results obtained, the items of the test were analyzed with respect to their difficulty and their biserial correlations with the appropriate subtests. On the basis of this analysis, the items of Forms 2 and 3 of the Officer Qualification Test were selected.

DESCRIPTION. It was decided that the Officer Qualification Test should consist of three parts—vocabulary, mechanical comprehen-

sion, and arithmetical reasoning. The choice of these three parts was made on both empirical and rational grounds. In the first place, independent verbal, mechanical, and arithmetical abilities are indicated in factorial analyses of mental ability. Secondly, tests of these abilities appear to be requisite for valid selection of candidates for naval officer training schools.

The vocabulary part of the Officer Qualification Test was chosen as a representative measure of verbal abilities, which are essential to success at officer candidate school. The ability to comprehend quickly and easily the material presented in texts and lectures was considered especially valuable at a time when the training courses at these schools were condensed and presented to the students at a greatly accelerated rate.

The mechanical comprehension part of the Officer Qualification Test was intended to measure some of the mechanical abilities which are an important requirement of naval officer candidates and which have become increasingly important as the Navy becomes more extensively mechanized. Much of the course material in officer training schools deals with problems which involve mechanical equipment and mechanical situations. As a result, success in officer training schools is probably dependent in considerable measure on the ability to comprehend the details of problems arising in such situations.

The arithmetical reasoning part of the Officer Qualification Test was intended to measure the candidate's ability to deal with quantitative problems. It was considered important to have an arithmetical part in the test for two reasons. First, training in mathematics is emphasized by the Navy for all officers. Second, thinking in quantitative terms is required both in officer training schools and in the performance of many shipboard duties.

Considerable differences in academic training in mathematics were to be expected in the candidates for naval commissions. It would have been an unfair penalty to those without extensive mathematical training to include in the quantitative section of the Officer Qualification Test items which required for their solution a knowledge of algebra, trigonometry, or college mathematics. Therefore, items were chosen for this subtest which could be solved by simple arithmetical operations. The main part of the task, however, was to choose the correct operations. The attempt was made to select items so that the time required for the actual computations would be only about one-fifth of the time needed for completing the test.

The Officer Qualification Test contains a total of one hundred items. The *Vocabulary* subtest consists of fifty "opposites" items. The task on each item is to select from a list of five the one word which most nearly means the opposite of the given stimulus word.

The candidate indicates the word selected on his answer sheet. This type of item is illustrated in the preceding chapter.

The *Mechanical Comprehension* [1] subtest consists of thirty items. Each contains a picture illustrating a mechanical situation and a question about the situation. Three answers are given to each item; the task is to choose the correct answer and indicate the choice on the answer sheet. It was intended that high scores on this test should indicate persons who can make reasonable common-sense judgments about a variety of mechanical situations. The preceding chapter contains an illustrative item of this type.

The *Arithmetical Reasoning* subtest contains twenty questions, and five alternative answers to each. The task is to decide, by performing the required arithmetical operations on the data given, which answer is correct. The choice is then to be indicated on the answer sheet. An illustration of this variety of item is also given in Chapter VI.

Immediately below is a tabular description of the subtests of the Officer Qualification Test.

Subtest	Number of Items	Time in Minutes [2]
Opposites	50	15
Mechanical Comprehension	30	15
Arithmetical Reasoning	20	25
Checking time		5
Total	100	60

[2] Only the total time limit is strictly observed. The time to be allotted to each of the three parts of the test is suggested in the instructions. The total time allowed includes five minutes to be used on items originally skipped.

USE OF THE OFFICER QUALIFICATION TEST. The Officer Qualification Test has been used in offices of naval officer procurement as an intellectual screening device. In addition to an applicant's score in the test, his educational, vocational and personal qualifications have been considered in the processing of his application (see Chapter II). Since the Officer Qualification Test has been used as a selection device, a critical or "cutting" score of 40 (Navy Standard Score) was adopted as the lower limit of eligibility of applicants, except in some cases where a useful experience was extensive enough to compensate for a lower score on the test.

[1] Many of the items in the Test of Mechanical Comprehension, Form BB, loaned by the Psychological Corporation, were used in the first experimental form, SE O-I. The later tests consisted of original items.

By the selection of a comparatively low cutting score it has been possible to eliminate from consideration only those candidates well below the mean, who for intellectual reasons would very probably not have succeeded in primary training. Since the need for officers was great and the majority of applicants were of college graduate caliber, it was not deemed necessary to set a high cutting score in the test.

Statistical Data for the Officer Qualification Test

The statistical data on the experimental and final forms of the Officer Qualification Test will be presented under the following topics: norms, means and standard deviations, reliability and intercorrelations. Except where otherwise indicated, the coefficients of reliability have been computed by the split-half method, corrected by the Spearman-Brown Prophecy Formula. Intercorrelations of test parts and validity coefficients have been computed by the product-moment method, and correlations between items and parts are biserial coefficients.

NORMS. The problem of constructing norms for the Officer Qualification Test was complicated by the fact that the test was to be used for officer populations from all sections of the country and there was reason to believe that samples drawn from different offices of naval officer procurement might not be equivalent. Added to this was the fact that men and women officers performed differently, especially on the Mechanical Comprehension subtest. It was decided therefore to prepare separate norms for men and women for Form 1 on the basis of the answer sheets of all tests given over selected periods at all offices.

MEANS AND STANDARD DEVIATIONS FOR THE OFFICERS' SELECTIVE EXAMINATION (SE O-1) AND OFFICER QUALIFICATION TEST, FORM 1. The first experimental form of an officer selection test (SE O-1) was put into use in the New York Office of Naval Officer Procurement. This test contained 115 items from which, after analysis, 100 were to be chosen for the Officer Qualification Test, Form 1.

Table 1-VII shows the means and standard deviations of the subtests and total tests of the experimental version (SE O-1), and of Form 1 for both New York and national samples.

The means of the subtests indicate a satisfactory difficulty, and the size of the standard deviations shows a desirable dispersion. The subtests, particularly Arithmetical Reasoning, were somewhat easier in the experimental than in the final form.

RELIABILITY COEFFICIENTS AND INTERCORRELATIONS. The reliability coefficients of the subtests and of the total tests, and the intercorrelations of the subtests of both forms as determined from three

TABLE 1-VII. Officer Qualification Test, SE O-1 and Form 1. Means and standard deviations (raw scores) on the subtests and total test. (N = 500 for each of the three samples.)

	SE O-1			Form 1				
	Number of Items	New York Sample		Number of Items	New York Sample		National Sample	
		M	σ		M	σ	M	σ
Subtest								
Opposites	60	32.28	11.91	50	27.77	10.83	25.73	10.46
Mechanical Comprehension	30	18.50	4.39	30	15.98	3.73	16.43	3.73
Arithmetical Reasoning	25	15.76	4.76	20	10.23	4.19	10.42	4.03
Total Test	115	66.54	16.84	100	53.99	14.82	52.49	13.93

samples, are shown in Table 2-VII. In all cases, the Opposites and Arithmetical Reasoning parts and the total test are satisfactorily reliable, while the Mechanical Comprehension part is not.[3] The

TABLE 2-VII. Officer Qualification Test, SE O-1 and Form 1. Reliability coefficients[1] and intercorrelations of subtests and total test. (N = 500 for each of the three samples.)

	Subtest							
	Opposites		Mechanical Comprehension		Arithmetical Reasoning		Total Test	
	N.Y.[2]	U.S.[3]	N.Y.	U.S.	N.Y.	U.S.	N.Y.	U.S.
SE O-1								
Opposites	**.94**							
Mechanical Comprehension	.28		**.73**					
Arithmetical Reasoning	.42		.56		**.87**			
Total Test	.90		.62		.72		**.93**	
FORM 1								
Opposites	**.93**	**.92**						
Mechanical Comprehension	.23	.16	**.64**	**.62**				
Arithmetical Reasoning	.42	.36	.47	.38	**.86**	**.86**		
Total Test	.91	.90	.55	.50	.70	.66	**.93**	**.90**

[1] Reliabilities were estimated by applying the Spearman-Brown Prophecy Formula to odd-even correlation coefficients. Reliability coefficients are shown in bold face type.
[2] New York samples.
[3] National sample.

relatively low reliability of this section was assumed to be due in part to its heterogeneous nature. The technique of split-half correla-

[3] In the research on the Navy's aptitude testing program, reliability coefficients from .80 to .89 have been considered satisfactory; those of .90 or higher have been considered very satisfactory; and those below .80 were considered unsatisfactory for tests described in this chapter.

tion, which was used in these computations, gives too low an estimate of the reliability of a heterogeneous test.

The intercorrelations of the subtests are also shown in Table 2-VII. These intercorrelations and the correlations between subtests and total score are somewhat lower for the Officer Qualification Test, Form 1, than for the Officers' Selective Examination. Opposites correlated highest, and Mechanical Comprehension lowest, with total score. As a consequence of the high correlation of the Opposites subtest with total score, this subtest received comparatively heavy weighting in the total test score. The reliability of the total test was due largely to this heavy weighting of the most reliable subtest in the total score.

MEANS AND STANDARD DEVIATIONS FOR THE OFFICER QUALIFICATION TEST, FORMS 2 AND 3. The preparation of Forms 2 and 3 of the Officer Qualification Test was based on an item analysis of the experimental form (O-2), which contained 274 items including those in Form 1. Form O-2 was given to about 400 men at a naval indoctrination school and to about 400 women at a women's reserve midshipmen's school. The subtests were printed in separate booklets and were given separately at both schools (over a three-day period at the indoctrination school, and a two-day period at the women's reserve midshipmen's school). About four hours were allowed for all of the tests. This time was considered sufficient to compensate for the large number of items (274) in the tests. On the basis of item analyses of the results of these two administrations of the experimental form (O-2), two new forms (Forms 2 and 3), which were equivalent in difficulty and item-subtest correlations were prepared. The attempt was made to select pairs of items, one for each form, in which the two items were equal in difficulty and item-subtest correlation. This procedure matched the two forms item by item rather than just in mean and standard deviation. No attempt was made, however, to make Forms 2 and 3 equivalent to Form 1.

Table 3-VII shows separately for men and women the means and standard deviations of the three forms of the Officer Qualification Test. It can be seen, first, that the new forms, 2 and 3, were equivalent with respect to their mean scores and standard deviations, but were somewhat easier than Form 1. Second, it is apparent that on Forms 2 and 3 the men and women performed differently. The Opposites part was easier for the women, while the Mechanical Comprehension was easier for the men.

Analysis of the items of the tests had also shown that with respect both to item-difficulty and to item-subtest correlation there were pronounced sex differences. The Opposites items were generally easier for the women, and the Mechanical Comprehension items

TABLE 3-VII. Officer Qualification Test, Forms 1, 2, and 3. Means and standard deviations (raw scores) on the subtests and total test for men and women

	OQT Form Number	Population	Subtest: Opposites M	Opposites σ	Mechanical Comprehension M	Mechanical Comprehension σ	Arithmetical Reasoning M	Arithmetical Reasoning σ	Total Test M	Total Test σ
Men	1	National Sample	26.0	10.4	16.5	3.6	10.5	4.0	52.9	14.1
	1	Indoctrination School—Class 6	26.7	9.8	15.3	3.4	11.0	3.2	53.0	12.5
	2	Indoctrination School—Class 7	27.9	9.8	16.3	4.5	12.4	3.4	56.6	13.5
	3	Indoctrination School—Class 7	28.2	9.9	16.2	4.2	12.4	3.4	56.8	13.5
Women	1	National Sample	26.3	9.8	12.4	3.0	8.7	3.3	47.4	12.2
	2	Reserve Midshipmen's School (WR)	31.6	8.5	12.5	3.3	11.3	3.2	55.5	10.9
	3	Reserve Midshipmen's School (WR)	32.2	8.6	13.2	3.0	11.5	3.2	56.9	10.8

easier for the men. The biserial correlation coefficients between item and subtest were generally lower for the women. In view of all these differences, separate norms were reported for the women.

RELIABILITIES AND INTERCORRELATIONS, OFFICER QUALIFICATION TEST, FORMS 2 AND 3. The reliability coefficients of the subtests and of the total test are shown in Table 4-VII. The coefficients of the new forms were estimated in two ways: first by correlating the scores on Form 2 versus Form 3 where the same persons, taking the experimental form (O-2), answered the items of both forms. Second, the reliabilities were estimated on the basis of the correlations of odd-

TABLE 4-VII. Officer Qualification Test, Forms 2 and 3. Reliability coefficients of subtests and total test

Subtest	Alternate-Form (Correlation between Forms 2 and 3): Indoctrination School	Alternate-Form (Correlation between Forms 2 and 3): Reserve Midshipmen's School (WR)	Odd-Even (Corrected by Spearman-Brown Prophecy Formula): National Sample Form 2	Odd-Even (Corrected by Spearman-Brown Prophecy Formula): National Sample Form 3
Opposites	.92	.90	.91	.90
Mechanical Comprehension	.75	.50	.74	.78
Arithmetical Reasoning	.76	.75	.83	.83
Total Test	.91	.87	.91	.91

numbered versus even-numbered items. By either method of estimation, the reliabilities of the total test and of the Opposites subtest were very satisfactory. The reliability of the Arithmetical Reasoning subtest as computed by the split-half method was considered acceptable, and the reliability of the Mechanical Comprehension subtest was not considered satisfactory. Because of this low reliability and the lack of information about the differential validities of the subtests at the time the decision had to be made, it was deemed advisable to report total scores only on the Officer Qualification Test instead of reporting separate scores for the subtests.

The reliability coefficients of all the subtests, particularly of Mechanical Comprehension, were higher for the men than for the women, as is shown in the alternate-form correlations.

The difference between the two types of estimate of reliability of the Arithmetical Reasoning subtest is rather considerable (.83 vs. .76). The higher estimate obtained from the split-half correlation indicates that the attempt to match the items of this subtest in Forms 2 and 3 was not completely successful, although the use of two different samples may account for some of the difference between the two coefficients.

The intercorrelations among the subtests of Forms 2 and 3 are shown in Tables 5-VII and 6-VII. The correlations in Table 5-VII were computed for the items of Form O-2 which were retained for Forms 2 and 3. Since the candidates at the indoctrination school and women's reserve midshipmen's school took, in effect, both forms, it was possible to correlate the parts of Form 2 with those of Form 3. It can be seen that, in general, the intercorrelations are low enough to indicate a considerable degree of independence in the subtests, and to indicate the possible usefulness of reporting separate subtest scores, if the subtests could have been made sufficiently reliable. There is, also, a tendency for the women's scores to intercorrelate lower than the men's.

Table 6-VII shows the intercorrelations based on national samples of men who took either Form 2 or Form 3. The intercorrelations based on the national samples of men are of about the same magnitude as those based on the indoctrination school sample. In both cases, the Opposites subtest correlates highest, and Mechanical Comprehension lowest, with the total test score.

Validity of the Officer Qualification Test

The administration of the Officer Qualification Test, Form O-2, at the indoctrination school and the women's reserve midshipmen's school, and of Form 1 to about 400 men at the indoctrination school,

TABLE 5-VII. Officer Qualification Test, Forms 2 and 3. Intercorrelations of subtests and total test in Indoctrination School (Men) and Reserve Midshipmen's School (WR)

		Subtest											
		Mechanical Comprehension				Arithmetical Reasoning				Total Test			
		Indoc. School		Reserve Midship. School (WR)		Indoc. School		Reserve Midship. School (WR)		Indoc. School		Reserve Midship. School (WR)	
Subtest	Form	2	3	2	3	2	3	2	3	2	3	2	3
Opposites													
Indoctrination School	2	.23	.18			.30	.34			.88	.82		
	3	.23	.22			.31	.35			.83	.89		
Reserve Midshipmen's	2			.14	.12			.19	.19			.88	.80
School (WR)	3			.13	.10			.20	.21			.80	.89
Mechanical Comprehension													
Indoctrination School	2					.43	.44			.61	.51		
	3					.41	.46			.49	.58		
Reserve Midshipmen's	2							.38	.40			.52	.36
School (WR)	3							.31	.34			.33	.46
Arithmetical Reasoning													
Indoctrination School	2									.62	.55		
	3									.59	.65		
Reserve Midshipmen's	2											.55	.46
School (WR)	3											.48	.56

TABLE 6-VII. Officer Qualification Test, Forms 2 and 3. Intercorrelations, means, and standard deviations (raw scores) of subtests and total test, national samples. (Form 2, N = 561; Form 3, N = 574)

	Subtest							
	Opposites		Mechanical Comprehension		Arithmetical Reasoning		Total Test	
Subtest	Form 2	Form 3	Form 2	Form 3	Form 2	Form 3	Form 2	Form 3
Opposites			.27	.21	.39	.37	.87	.87
Mechanical Comprehension					.53	.44	.65	.60
Arithmetical Reasoning							.70	.69
M	26.98	27.50	18.26	17.77	11.50	11.50	56.80	56.77
σ	10.05	10.04	4.75	4.74	3.89	4.03	14.60	14.36

made possible a study of the extent to which school grades could be predicted from scores on the test. Such a study of the validity of the test differs in several respects from the usual validity study, and might better be considered a pre-validation of the Officer Qualification Test. In the usual method of validating a test, the finished form is administered to a group *before* they enter the work or training against which the test is validated. In the present instance, the test was given to persons already in training. Also, it was given in experimental form, which differed from the finished test in that it contained a large number of items not used in the final Forms 2 and 3, was given with much larger time allowances, and in three sessions. The validity coefficients obtained under those circumstances may therefore be different from those which would be obtained under the usual circumstances.

The curricula at both schools were divided into a number of general courses, each of which consisted of several subjects. Detailed school grades were obtained for each of the persons tested, and these were correlated with test scores. The courses taught at the indoctrination school were seamanship, ordnance, and navigation. In addition to the course grades, an Officer's Aptitude Rating of each man was available. This rating was based on judgments made by instructors of the candidates' potential officer qualifications. At the women's reserve midshipmen's school, all women first took the basic indoctrination course and then went on either to advanced indoctrination or to communications. Since some of the courses at both schools consisted of a number of different subjects, separate grades on each subject were obtained whenever possible.

RELIABILITY OF SCHOOL GRADES. The course grades at both schools were fairly reliable (.75 to .92), with the exception of the communications course at the women's reserve midshipmen's school. The reliability coefficients for this course were attenuated by the fact that grades in very dissimilar subjects were correlated to determine the reliability of the course grades.

CORRELATION WITH SCHOOL GRADES.[4] The Officer Qualification Test correlated with the final average grade at the indoctrination school to a serviceable degree (.50). A zero-order coefficient of correlation of this magnitude indicates a useful validity for the test in the selection of candidates for indoctrination school.

The correlation coefficients of Officer Qualification Test scores with the final weighted average grade at the women's reserve midshipmen's school were somewhat lower, ranging from .33 to .47. The lower validity of the test at this school was due partly to the lower reliabilities both of the subtests of the Officer Qualification Test and of the grades. Because of the relatively low validity of this test, the use of a test with somewhat different content was considered for use in the selection of WAVES officers.

MULTIPLE CORRELATIONS OF OFFICER QUALIFICATION TEST SUBTESTS WITH GRADES. There are several ways of making a test more valid for predicting a particular criterion. Among these are:

1. To increase the reliability of the test.
2. To change the weighting of the parts of the test.
3. To change the content of the test.

An increase in the reliability of a test in order to improve its validity is desirable and, in the present instance, was possible. In order to obtain a substantial increase in the validity of the Officer Qualification Test, it would need to be considerably longer. Furthermore, a good deal of change would have to be made in the Mechanical Comprehension part, by lengthening and purifying this subtest so as to make it more homogeneous. In view of the variety of different mechanical situations which naval officers are called upon to face, however, it was deemed inadvisable to eliminate any of the physical principles illustrated by the items in the Mechanical Comprehension part of the Officer Qualification Test. But only one hour was available for testing at offices of naval officer procurement and the length of the test was consequently limited. Because of these factors extensive revision of the test did not seem to be feasible.

The maximum improvement which could be expected from altering the weighting of the subtests of the Officer Qualification Test

[4] For a more detailed discussion of the validity of the Officer Qualification Test, see Chapter X.

can be ascertained by computing the multiple correlations of the three part-scores with the criteria and comparing these coefficients with the correlations between the scores on the total test and the criteria. The multiple correlations were higher than the correlations of total scores with final grades by amounts ranging from .03 to .07. Increments of this size are desirable, but in view of the thousands of men tested in many centers in a comparatively short time, and of the risks of computational errors in the weighting process, it was decided that predictions of success in primary training school might as well be based on the total test score. For purposes other than the selection of officer candidates for primary training it was thought best to devise a new test with additional content.

Experimental Studies of the Officer Qualification Test

During the development of the three forms of the Officer Qualification Test, and in certain cases after the tests were put into nation-wide use, studies were performed in order to obtain additional information.

Effect of a Time-limit on the Officers' Selective Examination (SE O-1). On the first day on which the Officers' Selective Examination was given at the New York Office of Naval Officer Procurement, the applicants were allowed an unlimited time for its completion. According to the report from that office, some applicants took as long as three and one-half hours on the test. Subsequently, a time limit of one and one-half hours was imposed. The means and standard deviations of the three subtests and of the total score under both conditions of time allowance were compared. The results are shown in Table 7-vii.

Table 7-vii. Officers' Selective Examination (SE O-1). Comparison of means and standard deviations (raw scores) for limited vs. unlimited [1] time

	Time	N	M	σ
Subtest				
Opposites	1½ hrs.	500	32.28	11.91
	Unlimited	92	32.11	9.63
Mechanical Comprehension	1½ hrs.	500	18.50	4.39
	Unlimited	92	19.04	4.17
Arithmetical Reasoning	1½ hrs.	500	15.76	4.76
	Unlimited	92	17.51	3.91
Total Test	1½ hrs.	500	66.54	16.84
	Unlimited	92	68.74	13.80

[1] Values for unlimited time were calculated from grouped data; hence the sum of the means for the three parts does not equal the mean for total score.

The effect of the time-limit, as shown in Table 7-VII, was slight and was confined mainly to the last part of the test, Arithmetical Reasoning. On the basis of the small effect of the time-limit on the standard deviations of the scores, it was concluded that the test functioned as a power test with the one and one-half hour time-limit.

STABILITY OF ITEMS IN OFFICER QUALIFICATION TEST, FORM 1 RETAINED FROM OFFICERS' SELECTIVE EXAMINATION (SE O-1). The experimental examination (SE O-1) was given at the New York Office of Naval Officer Procurement, and on the basis of an item-analysis of the results, Officer Qualification Test, Form 1, was devised. It was decided to determine the stability of the items of the experimental examination which were retained for Form 1, when administered to a new sample taken from the same population. For this purpose, a sample of answer sheets of Form 1, administered at the New York office, was compared with those obtained in the administration of the experimental examination at the same office. The items common to both tests were compared with respect to difficulty (p) and item-subtest correlation (r_{bis}).

Statistical means were used to evaluate the significance of changes in the difficulty on item-subtest correlation of all items from the first to the second administrations. The chi-square test was used to determine whether an item changed significantly in difficulty from one test to another. The difference in r_{bis} on the two administrations was compared with its standard error, and the size of the ratio was used to judge whether or not the change was significant.[5]

In both difficulty value and biserial correlation coefficient the changes in the values of the items were small and for the most part not significant. It was concluded that the items retained for Officer Qualification Test, Form 1, were satisfactorily stable.

Data were available on 74 items common to both tests. Twelve of these items changed in difficulty by amounts significant at the one per cent level (five opposites, four mechanical comprehension, three arithmetical reasoning), ten being more difficult in Form 1, as might be expected from the more severe time-limit. Of the 74 biserial correlation coefficients between items and their subtests, eight changed significantly at the one per cent level, with the general trend toward slightly lower correlations in Form 1.

COMPARISON OF NEW YORK AND NATIONAL SAMPLES, OFFICER QUALIFICATION TEST, FORM 1. The preparation of the Officer Qualification Test, Form 1, which was intended for use at all offices of naval officer procurement, was based on an analysis of data from a single office, New York. It was important, therefore, to determine whether the items functioned in the same way in the New York and

[5] See Appendix E-1 for a detailed discussion of the techniques used.

the national populations. An analysis was made of a random sample of 500 answer sheets of Form 1 submitted by all offices of naval officer procurement throughout the United States, and the results were compared with those obtained at the New York office. The comparisons were made on the basis of item difficulty and item-subtest correlation.

On the basis of the analyses it was concluded that the generally risky procedure of constructing a test for administration to a population on the basis of an analysis of the results obtained from a single sample of the population was successful in the present case. Only eight of the 100 difficulty values and eleven of the biserial correlations were significantly different at the one per cent level in the two samples. The Opposites items appeared slightly easier, and Mechanical Comprehension slightly more difficult for the New York sample. In general, the item-subtest correlation coefficients of the Opposites items were very slightly higher for the New York sample.

TABLE 8-VII. Officer Qualification Test, Form O-2. Comparison of mean scores and standard deviations (raw scores) for samples of men (in Indoctrination School), and women (in a Reserve Midshipmen's School [WR])

Subtest	Mean		Standard Deviation	
	Men	Women	Men	Women
Opposites	71.0	78.9	20.3	17.9
Mechanical Comprehension	53.9	46.2	9.9	7.3
Arithmetical Reasoning	32.8	30.7	7.3	6.9

COMPARISON OF MEN AND WOMEN ON THE OFFICER QUALIFICATION TEST, FORM O-2. The experimental test (Form O-2) from which Forms 2 and 3 of the Officer Qualification Test were constructed was given as has been indicated to a number of students at an indoctrination school for men and at a women's reserve midshipmen's school. The means and standard deviations of the tests, and the difficulty values and biserial correlation coefficients of the items were compared for the two groups in order to determine the relative performance of men and women on the test.

The means and standard deviations of the scores of the two groups are shown in Table 8-VII. The women's means were significantly higher (1% level) in Opposites and Arithmetical Reasoning, and significantly lower in Mechanical Comprehension.

The item analyses showed that about one-half of the Opposites items were significantly easier for the women, while about one-half of the Mechanical Comprehension items were significantly easier for the men. The analysis of the item-subtest correlation coefficients

also showed sex differences, with the coefficients generally lower for the women, particularly in Mechanical Comprehension. In view of the differences in mean score and standard deviation on the tests, and of the differences in difficulty and biserial correlation of the items, it is clear that men and women indoctrinees did not perform in the same way on the Officer Qualification Test.

Comparison of Indoctrination School and National Samples, Forms 2 and 3. The items of Forms 2 and 3 of the Officer Qualification Test were chosen on the basis of an analysis of data from administration of the items to already selected officer candidates at an indoctrination school. A study was undertaken to determine how the items performed for a nationwide sample from various offices of naval officer procurement.

By a random selection, a sample of 561 candidates who had taken Form 2 and 573 who had taken Form 3 were drawn from the answer sheets submitted by all offices of naval officer procurement during a two-week period. To determine whether the items were satisfactory for nationwide use, the indoctrination school data were compared with the national data.

Both in difficulty and in item-subtest correlation, the Opposites and Arithmetical Reasoning items were stable, but the Mechanical Comprehension items showed a number of significant changes. In general, the items of the latter subtest were easier and the correlation coefficients higher for the national sample. In view of this fact it was concluded that the preferable procedure would have been to construct Forms 2 and 3 on the basis of analyses of data from a nationwide sample of officer candidates.

Correlations of Officer Qualification Test Scores with Age. Studies were made at the indoctrination school and women's reserve midshipmen's school to determine the relation of Officer Qualification Test scores to the age of student officers in these primary training schools. The results are shown in Table 9-vii. The correlation coefficients of test scores with age were low at both schools, and were mostly not significantly different from zero. There is an indication, however, that the older students tended to do somewhat better on the Officer Qualification Test, particularly the women; this was most noticeable on the Opposites subtest.

Evaluation. The Officer Qualification Test functioned satisfactorily as an aid in deciding whether or not an applicant was potential officer material. It correlated quite well with subsequent grades in primary training schools and it gave evidence of being a good measure of ability to profit from training. It therefore seems likely that the men and women who did not receive commissions at least in part because they had poor test scores would not have

been able to learn about and adapt to the Navy rapidly enough to make a real contribution to the Service.

The Officer Classification Test

The Officer Classification Test (OCT) was constructed in order to provide differential information for the classification of indoctrinees

TABLE 9-VII. Officer Qualification Test, Forms 1, 2, and 3. Correlation coefficients of scores with age in years of students in Indoctrination School and Reserve Midshipmen's School (WR)

	Indoctrination School Class 6, N = 360 Class 7, N = 403			Reserve Midshipmen's School (WR) Company 1, N = 152 Company 2, N = 160 Company 6, N = 115		
	Form	Class	r	Form	Company	r
Subtest						
Opposites	1	6	.20			
	2	7	.10 [1]	2	1	.36
					2	.25
					6	.31
	3	7	.10 [1]	3	1	.34
					2	.25
					6	.27
Mechanical Comprehension	1	6	.04 [1]			
	2	7	.03 [1]	2	1	.02 [1]
					2	.05 [1]
					6	—.05 [1]
	3	7	.05 [1]	3	1	.01 [1]
					2	.02 [1]
					6	—.26
Arithmetical Reasoning	1	6	.12 [1]			
	2	7	—.03 [1]	2	1	.15 [1]
					2	.06 [1]
					6	—.22
	3	7	—.06 [1]	3	1	.14 [1]
					2	.17 [1]
					6	—.26
Total Test	1	6	.20			
	2	7	.07 [1]	2	1	.32
					2	.23
					6	.18 [1]
	3	7	.07 [1]	3	1	.32
					2	.25
					6	.06 [1]
Mean Age in Years		6	29.73		1	25.12
		7	29.74		2	25.03
					6	30.10
Standard Deviation (Age in Years)		6	4.30		1	3.74
		7	4.17		2	3.21
					6	6.35

[1] Coefficients are not significantly different from zero at the 1% level.

or midshipmen to facilitate their assignment to further training or to specific ship or shore billets. Additional impetus was given to this project by a request from the Amphibious Training Command, Atlantic Fleet, for an instrument which could be used in the distribution to billets of officers in the amphibious forces. The content of the Officer Classification Test was determined with several considerations in mind. First were the results of factorial studies of mental organization. Second was the knowledge of what tests had been successful in the Navy on the enlisted level. Third was the experience gained with the Officer Qualification Test and other predecessor tests used with officers, namely, the Pre-Radar Battery (described in Chapter VIII), and the Officer Mechanical Aptitude Test, experimental form.

The tests of the Pre-Radar Battery and the Officer Mechanical Aptitude Test were to be used, however, in selection for specific schools or billets, rather than for classification or distribution of personnel to a wide variety of schools or billets. The problem in selection is to determine whether or not an individual is qualified to serve as a member of an organization. In classification, the problem is to determine, from the various aptitudes indicated by the tests, in which of several kinds of training or work an individual is most likely to succeed. A classification battery must therefore provide evidence of the varying levels of ability of an individual in different kinds of work. For this purpose, part-scores are recorded separately and the scores on the total test are either a series of numbers or a graphic "profile." Because decisions are based on part-scores and on the differences between part-scores, such a battery requires higher reliability and validity for each part than is necessary in a selection test in which no part-scores are reported.

The Officer Qualification Test did not provide differential information regarding candidates, since only total scores were reported. It was partly because such differential information regarding various abilities was needed, that the decision was made to use a relatively untested experimental form of the Officer Classification Test while evidence of its suitability accumulated as a basis for its revision.

Until the end of the war, Form X-1 of the Officer Classification Test was used for the purpose of classifying officers for assignment beyond their indoctrination or midshipman training. During the use of Form X-1, the need for various modifications became apparent, and these changes were incorporated in the new Officer Classification Battery. The end of the war found this battery in the process of development.

DESCRIPTION. The Officer Classification Test consists of four sections: Verbal, Mechanical, Mathematical, and Spatial. The Verbal Section contains 60 opposites items of the kind used in the Officer

Qualification Test. Since the opposites part of the Officer Qualification Test was very reliable, and a larger number of these items was included in the new test, the Verbal Section of the Officer Classification Test was expected to be highly reliable.

The Mechanical Section of the Officer Classification Test has two subsections. The first is Mechanical Comprehension, with items of the same kind as in the Officer Qualification Test, but in greater number (45) in order to obtain a satisfactory reliability. The second subsection, Electrical and Mechanical Information, intended to measure additional aspects of mechanical abilities, contains 45 items of the kind described in the previous chapter for the Mechanical Knowledge Test for enlisted personnel.

The Mathematical Section of the Officer Classification Test contains 45 items of the same kind as the arithmetical reasoning items

DESCRIPTION OF THE OFFICER CLASSIFICATION TEST

	Number of Items		Time Limits in Minutes	
	Part	Total	Part	Total
Part I. Verbal Section		60		20
Opposites	60		20	
Part II. Mechanical Section		90		30
Mechanical Comprehension	45		20	
Electrical and Mechanical Information	45		10	
Part III. Mathematical Section		45		45
Mathematical Problems	45		45	
Part IV. Spatial Section		60		35
Block Assembly	30		15	
Rotation of Solid Figures	30		20	

in the Officer Qualification Test. The reliability of this part of the Officer Classification Test was expected to be high because of the nature of the items and the number used. The Spatial Section of the Officer Classification Test consists of two subsections which require the ability to visualize three-dimensional spatial relations. These spatial abilities are indicated as independent of other mental abilities in factor studies of mental tests.

Uses of the Officer Classification Test

The Officer Classification Test was developed in response to a need for a classification instrument at reserve midshipmen's schools, indoctrination schools, and certain operational training commands.

Classification officers at these installations were faced with the task of recommending student officers for advanced training or billet assignments. These school and billet assignments could be roughly divided into three major categories: highly technical, e.g. radar officer; semi-technical, e.g. communications officer; non-technical or administrative, e.g. motor torpedo boat skipper. Tentative cutting scores on appropriate sections of the test were established for the various schools and billets falling in these categories, and assignments were made, in part at least, on the basis of test scores. The procedures followed by the classification officers are described in Chapter II; the effectiveness of the Officer Classification Test in predicting success of student officers in various types of training is discussed in Chapter XI. Validity data from some of the studies are presented in this chapter.

Statistical Data on the Officer Classification Test

NORMS. The problem of constructing norms for the Officer Classification Test was complicated by the fact that the test was intended for use with two officer populations, indoctrinees and midshipmen. There was no concrete evidence of the equivalence of these two groups with respect to test score. Owing to the differences in the average age of the two groups, it was anticipated that the midshipmen might be slightly superior in mathematical and spatial ability and slightly inferior in verbal and mechanical ability. Previous studies of the Officer Qualification Test and of the Basic Test Battery for enlisted men had indicated that mechanical and verbal test scores correlated positively, while mathematical and spatial test scores correlated negatively with age, all to a slight extent. In view of this possibility, it was decided that the norm group should contain a sample from each population, and that the two should be combined if the data warranted the combination.

Accordingly, the test was administered to 500 midshipmen at a reserve midshipmen's school and to 500 indoctrinees at an indoctrination school. Slight differences were found between the samples; these were in the expected direction, but were so small as to be of no practical significance. The samples were therefore combined to form a norm group of 1,000. Four separate distributions were made for the four parts of the test, and four Navy Standard Score (NSS) scales were derived from the respective means and standard deviations. The Navy Standard Score is based on a distribution in which the mean is assigned a scale value of 50 and the standard deviation a value of 10. When used in the classification process the Officer Classification Test scores are expressed in this form.

MEANS, STANDARD DEVIATIONS, RELIABILITY COEFFICIENTS, AND INTERCORRELATIONS. Table 10-VII summarizes the data obtained from a sample of 250 of the norm group to which the Officer Classification Test, Form X-1 was administered. The table contains the means, standard deviations, reliability coefficients, and intercorrelations of the sections and subtests of the Officer Classification Test. An examination of the table leads to the following observations.

1. The locations of the means indicate the possibility of a slight skewness in the distributions.

2. The intercorrelations among the sections of the test are suitably low for a classification test. The independence of the subtests, illustrated particularly by the correlations of .08 between the Verbal and Mechanical Sections, and .02 between the Verbal and Spatial Sec-

TABLE 10-VII. Officer Classification Test, Form X-1. Means, standard deviations, intercorrelations, and reliability coefficients. (N = 250, proportionate sample of norm group)

Section of Test	M [2]	σ	Reliability Coefficients [1] Part I	Part II	Part IIA	Part IIB	Part III	Part IV	Part IVA	Part IVB
Part I, Verbal	32.72	12.60	**.92**							
Part II, Mechanical	53.24	11.01	.08	**.83**						
A. Comprehension	26.88	6.23	.12	.85	**.75**					
B. Information	26.28	6.48	.03	.87	.52	**.74**				
Part III, Mathematical	23.77	6.79	.24	.38	.45	.25	**.78**			
Part IV, Spatial	37.12	8.56	.02	.34	.43	.18	.38	**.81**		
A. Block Assembly	20.64	3.97	.04	.31	.40	.17	.37	.84	**.55**	
B. Rotation of Solid Figures	16.51	5.50	.00	.32	.41	.18	.36	.93	.62	**.77**

[1] Reliability coefficients were computed by the Kuder-Richardson Formula, No. 21. They are shown in bold-face type.

[2] Means and standard deviations expressed in raw score form.

tions, indicates the feasibility of using part-scores in classifying officers.

3. The reliability of the Mathematical Section (.78) is not as high as is usually considered necessary for a classification test. The reliability of the Mechanical Section (.83) is satisfactory but not high.

VALIDITY. There are three slightly different types of information about the validity of the Officer Classification Test. These will only be sampled in what follows, since they are covered in connection with specific schools and training programs in Chapter XI of this volume.

First, there is the type of evidence which is exhibited by plotting the test profiles of certain groups successful in training for a given type of billet by comparison with the average officer in the norm

group. Figure 1-VII shows three profiles of this sort, in which the units of measurement on the ordinate are Navy Standard Scores (NSS). It will be observed that each of these profiles differs from the others and from the Navy norm in a marked fashion. Especially noteworthy are the following facts:

1. The pre-radar curriculum bristles with mathematics and theoretical physics. The average officer who completes the pre-radar training is superior to 93% of the officers in the norm group on the Mathematics Section of the Officer Classification Test.

2. The curriculum in diesel engineering places a heavy emphasis upon mechanical facility, and the average officer in this training

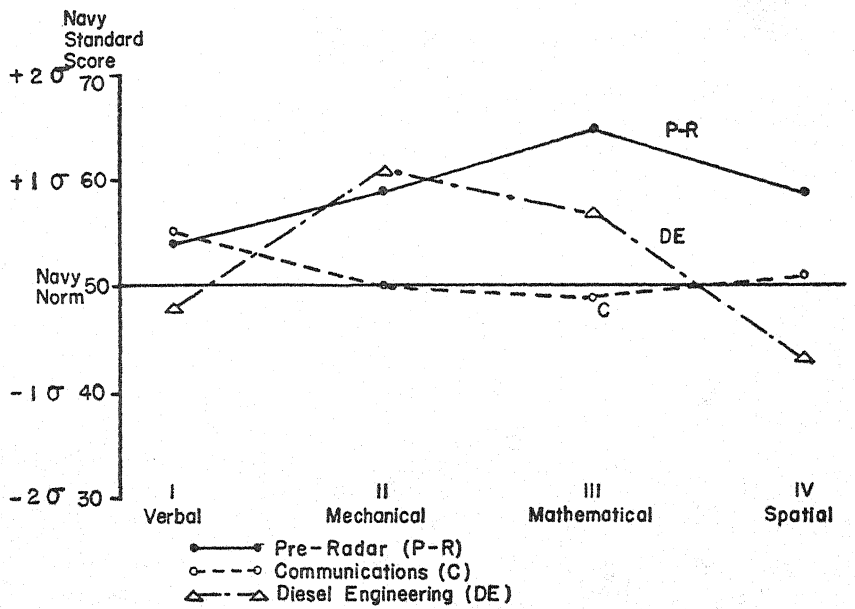


Figure 1-VII. Profiles of mean scores of three groups on all sections of the Officer Classification Test.

scores at the 86th percentile of the norm group on the Mechanical Section of the Officer Classification Test.

3. The communications group, in agreement with expectation, is most select with respect to verbal ability. The average graduate of communications school scored at the 69th percentile of the norm group on the Verbal Section of the test.

A second type of evidence of validity may be derived from the correlations of the sections and subtests of the Officer Classification Test with course grades in an advanced school (Submarine). Table 11-VII contains data of this sort for the Officer Submarine School. It

will be observed that the Mathematics and Block Assembly scores show the best average correlations with grades in the several courses. It will also be noted that the grades in engineering are the most predictable ones given at the Officer Submarine School. Accordingly, the correlation of 0.48 between the Mathematics Section and the engineering grades is the highest in the table. A validity coefficient of this magnitude seems quite satisfactory for this type of training program.

A closely related bit of validation may be gained from the correlations of the sections and subtests of the Officer Classification Test with grade average at several advanced officer schools. Table 1-XI (Chapter XI) contains data of this type. In it, the Mechanical and Mathematical Sections of the test appear to be considerably superior to the Verbal and Spatial Sections. This fact is to some extent a

TABLE 11-VII. Officer Classification Test. Correlation coefficients between test scores and course grades at the Officer Submarine School. (N = 293)

Section of Test	Correlation Coefficient with Course Grades					
	Torpedoes	Communi-cations	Sub-marines	Engi-neering	General	Average
Part I, Verbal	.15	.25	.17	.10	.27	.19
Part II, Mechanical						
A. Comprehension	.23	.23	.23	.39	.21	.26
B. Information	.18	.14	.22	.34	.01	.18
Part III, Mathematical	.30	.36	.35	.48	.26	.35
Part IV, Spatial						
A. Block Assembly	.28	.30	.33	.35	.23	.30
B. Rotation of Solid Figures	.17	.15	.23	.22	.18	.19
Average of r's	.22	.24	.26	.31	.19	
Total Test	.32	.36	.39	.43	.30	.36

reflection of the inherent merits of the tests involved. But it seems probable that it is also a commentary upon the nature of the abilities required for success in the Navy. A knowledge of how to operate or repair apparatus of great mechanical complexity is undoubtedly at a premium in the naval service; and the ability to handle quantitative problems, from fire control to navigation, is at a comparable premium. The sections mentioned are, therefore, more valid at least in part because they demand abilities highly relevant to the duties of line officers.

REVISIONS. Since the form of the Officer Classification Test under discussion was actually an untested experimental form, it was anticipated that alterations and changes in content would ultimately be necessary. The changes made and the reasons for making them are as follows:

First, analogies items were substituted for the opposites items of the Verbal Section. This was done because of the previously noted unsatisfactory validity of the opposites items and because, at the enlisted level, analogies tests had shown considerably larger validity coefficients than the opposites tests.

Second, five-choice mechanical comprehension items were substituted for the three-choice mechanical comprehension and mechanical-electrical information items of Part 2. This was done because the comprehension items were superior in validity to the information items and because it was felt that the five-choice items might be more appropriate in difficulty and more reliable then the three-choice variety.

Third, the Mathematics Section was lengthened and made more homogeneous in content. This was done because of its unsatisfactory reliability (.78) and because of evidence which indicated that a number of items involving geometric orientations were not measuring the same thing as the rest of the test, with a resulting attenuation of the reliability.

Fourth, a Relative Movement Test was developed and incorporated in the battery. This was done since such a test, as part of a specialized battery, had predicted success very well in such schools as tactical radar, fighter director, and sonar; and because it did not correlate highly with other tests already included in the Officer Classification Test. The items consist of verbal statements about the maneuvers of one or more ships. The testee is to visualize the problem, and to answer in terms of a bearing or an approximate speed or distance.

Fifth, the Spatial Section was kept intact with only minor modifications because the evidence as to its worth was inconclusive. It predicted very well in a few schools, but it had a generally low validity. Also, the data as to the effectiveness of the Block Assembly Subtest as compared with the Block Rotation Subtest were inconsistent. One subtest did better in one school and the other in another. It was thus felt that this test should be retained for further evaluation.

Sixth, an effort was made to increase slightly the difficulty of the test by small reductions of the time limits. This was done because most of the distributions obtained on the Officer Classification Test were very slightly negatively skewed.

Finally, the Officer Classification Test was recognized to be more properly an Officer Classification Battery and the designation and format were changed accordingly. Each "section" was put in a separate booklet to facilitate the use or omission of individual tests as desired.

The Officer Classification Battery is composed as follows:

I	Verbal Reasoning Test	75 five-choice analogies items
II	Mechanical Comprehension Test	48 five-choice mechanical comprehension items
III	Mathematics Test	50 five-choice mathematics items
IV	Relative Movement Test	50 four-choice relative movement items
V	Spatial Test	
	A. Block Assembly	30 four-choice block assembly items
	B. Block Rotation	30 five-choice block rotation items

EVALUATION. In looking at the place of the Officer Classification Test in the testing programs of the Navy, the following summary statements concerning it seem justifiable: The Officer Classification Test was useful in the distribution of indoctrinees and midshipmen to billets. Evidence indicates that it did differentially predict potential school success, even within the very select groups referred for advanced officer training. Officer Classification Test scores were undoubtedly of considerable value to classification officers in determining the broad general area within which each individual might be expected to perform most efficiently.

College Qualifying Tests (C-Tests)

On April 2, 1943, the first Army-Navy College Qualifying Test (Test C-1) was given in over 14,000 educational centers throughout the United States. Upon the score made by each candidate on this test depended his eligibility for special instruction in the college training program of the Army or the Navy. The purpose of the test was to select from all eligible candidates of high school education or its equivalent, those men who would most likely absorb successfully the training to be given them, and who would later be officers of the Army or Navy. In this chapter, the discussion of the College Qualifying Tests will be limited to Test C-1, which was administered to more than 300,000 candidates for college training.

DESCRIPTION. The C-1 test included 150 questions of the multiple-choice type, for the completion of which the candidates were allowed two hours, divided as shown in tabular form below. There were four principal sections in the test, consisting respectively of verbal, scientific, reading, and mathematical problems.

Verbal Section. The verbal items were of three kinds: 30 opposites, 15 analogies, and 15 double definitions. Opposites and analogies

items have been illustrated in the preceding chapter. An example of double definitions follows:

> ..... is an opinion held in opposition to the established .....: (1) belief .. doctrine (2) stubbornness .. doctrine (3) heresy .. doctrine (4) heresy .. government (5) fascism .. government.

Science Section. The scientific section consisted of 40 problems of a common-sense physics type. The technical training required to answer these questions was not great. For the most part intelligent scientific interest and alert observation would be as valuable as academic training. An example of this type of item follows:

> The structure of an atom is analogous to that of (a) earth (b) the solar system (c) the human body (d) a dynamo.

Reading Section. The third section of the test consisted of four paragraphs of about 200 words each. Five questions followed each paragraph, and five answers were suggested to each question. The candidate was to choose the best answer. In this section the candidate's ability to understand and answer questions about relatively difficult material was under test. This ability is, of course, basic to success in college. Three of the paragraphs dealt with economic and historical material, while the fourth was on biology.

Mathematics Section. The fourth section of the test consisted of 30 questions designed to test the candidate's facility with numbers and required a background of elementary algebra and geometry.

The 150 items which were selected for the final test were chosen from a large number which were analyzed with respect to difficulty and item-subtest correlation as explained in Appendix E-1.

A summary description of the test is presented below.

Section of Test	Number of Items	Time Limit in Minutes
I. Verbal		30
Part 1, Opposites	30	
Part 2, Analogies	15	
Part 3, Double Definitions	15	
II. Scientific	40	30
III. Reading	20	25
IV. Mathematics	30	35
	150	120

MEANS, STANDARD DEVIATIONS, RELIABILITIES, INTERCORRELATIONS. Norms for the C-1 test were prepared from the answer sheets of a representative sample of 1,500 cases drawn from the population of men from 17 to 21 years of age, inclusive. Evidence of the representativeness of the sample is derived from the fact that the mean and standard deviation of the scores on the test differed by only very little between the sample and the total group of the same range of age. The test was scored as a whole when used for selection, but the means, standard deviations, reliabilities and intercorrelations of the sections, as computed from the sample of 1,500 cases, are shown in Table 12-VII.

The reliabilities of the Verbal Section and total test are seen to be satisfactory. The Reading Test, which had the lowest reliability

TABLE 12-VII. College Qualifying Test (C-1). Means, standard deviations, reliability coefficients, and intercorrelations of sections and total test. (N = 1500)

	Part of Test				
	Verbal	Science	Reading	Mathematics	Total Test
Part					
Verbal	**.93** [1]	.59	.75	.65	.92
Science		**.85**	.57	.66	.81
Reading			**.76**	.63	.83
Mathematics				**.87**	.83
Total Test					**.96**
Mean Score [2]	25.4	20.5	12.1	15.1	73.1
Standard Deviation	11.9	7.1	4.3	5.5	24.8

[1] The figures in bold-face type are reliability coefficients calculated by the Spearman-Brown formula from the correlation of odd- and even-numbered items.

[2] Means and standard deviations are expressed in raw scores, i.e., number of items correct.

(.76), contained only 20 items. The mean of each section is located between 42% and 61% of the maximum possible score, and the standard deviations are adequately large.

The intercorrelations among the sections of the test are sufficiently low, with one exception, to indicate that the subtests measured different abilities. The exception is the correlation between the Verbal and Reading Sections, .75, which would be expected to be high in view of the fact that the Reading Test is highly loaded with a verbal component of intelligence. It is noteworthy that the Science Section does not correlate much higher with Mathematics than with the Verbal and Reading Sections, and that Mathematics correlates equally with all of the other sections.

USE OF THE TEST IN SELECTION. Since the Navy accepted men for its college training program on the basis of a quota proportional to the population of each state, the critical test score for acceptance varied from 75 to 106, depending on the number to be accepted from each state. After receipt of the record sheets of the men whose scores were within the eligible range, the Navy notified those men of their success on the test. In addition, the records of all men with scores of 75 or more were sent to the offices of naval officer procurement and the district quotas were adjusted so that the Navy could avail itself of the candidates with good ability who would otherwise have been eliminated by the exceptionally high critical scores in their districts. In this way all men who made a score of 75 or better, regardless of their districts, actually received consideration for the college training program.

The success of the C-test in the selection of candidates for the college training program is discussed in Chapter X, together with the evidence on which the evaluation of the test is based.

Usefulness of Basic Tests for Officer Personnel

The tests discussed in this chapter have served a useful purpose in the selection and classification of officer personnel. It will be shown in Chapter VIII that there was little need for special aptitude tests in classification after the Officer Classification Test came into general use. Success in the type of training undertaken by naval officers can be predicted best by tests of verbal, mechanical, and mathematical ability; spatial tests seem to have less applicability. In considering an officer's qualifications for various types of training or billet assignments, it is almost always helpful to have information concerning the abilities measured by the Officer Classification Test. It seems desirable, therefore, to have these scores for all officers in the Navy. The new Officer Classification Battery will have the additional advantage of providing a score on the Relative Movement Test which was found to predict success very well in certain types of training. With these test scores it should be possible to predict with a reasonable degree of success how well most officers will succeed in advanced training and subsequent billet assignments.

CHAPTER VIII

SPECIAL APTITUDE TESTS

THE basic selection test batteries for officer and enlisted personnel described in Chapters VI and VII were developed to measure certain of those qualifications which are required for success in a wide variety of naval training schools, Navy jobs, and shipboard billets. Partly because of the highly specialized character of the abilities required for success in certain of these programs, and partly because the basic test batteries were not developed early enough, special aptitude tests were constructed.

The Radio Code Test used in the Navy is illustrative of the type of test developed to measure a specialized type of aptitude. Attrition in radio schools has always been higher than in most other types of training for enlisted personnel, and the original tests of the basic battery were found to be relatively poor predictors of ability to learn the code. The Radio Code Test—Speed of Response was accordingly developed to assist in screening out recruits with low code learning ability in order to reduce the failure rate in radio schools.

The CIC (Combat Information Center) Aptitude Test is illustrative of tests developed to fill a selection need before a basic comprehensive test battery had been developed. At the time that the tactical radar school for officers was established, the Officer Qualification Test was the only psychological test on the officer level standardized for use in the Navy. It was believed that the latter test alone would be inadequate for the purpose of selecting candidates for tactical radar school since it apparently did not measure certain aptitudes required for success in this type of training. As a result the special CIC Aptitude Test was developed. After the Officer Classification Test was developed, it was found that prediction of success in tactical radar training could be made almost as satisfactorily with the Mathematics Section of this test as with the CIC Aptitude Test. When the Officer Classification Battery was developed in 1945 the Relative Movement Test, which is Part 3 of the CIC Aptitude Test, was incorporated as a basic part of the battery. Modification of personnel policies following the surrender of Japan has, however, prevented putting into operation this battery of officer tests to determine whether it could completely replace the use of special aptitude tests on the officer level.

The special aptitude tests used by the Bureau of Naval Personnel can conveniently be divided into two groups. Those listed below

under Group I have been widely used in selection and classification of naval personnel and their usefulness in predicting success has been extensively studied. The tests listed in Group II have, in general, received rather limited use and as a result have not been studied as intensively as those in Group I. The Sonar Pitch Memory Test is placed in Group II because studies of this test have been made by the National Defense Research Committee rather than by the Bureau of Naval Personnel.

Group I Tests

THE CIC (COMBAT INFORMATION CENTER) APTITUDE TEST. When the plans for opening the tactical radar training school were completed in late 1943, the Test and Research Section was asked to develop a test for use in selecting candidates for this school from graduates of indoctrination and reserve midshipmen's schools. Analysis of the job of the tactical radar officer and of the tactical radar curriculum indicated that the following abilities would be required: (1) quickness of judgment, (2) speed and accuracy of plotting, and (3) quick visualization of the tactical situation. In addition it was believed that a high level of mental ability (to be measured by the Officer Qualification Test) would be necessary in order to comprehend, coordinate, and administer the complex activities of a combat information center. The result was the preparation of the Tactical Radar Aptitude Test Battery consisting of four subtests: Polar-Grid Coordinate, Ratio Estimation, Coordinate Reading, and Relative Movement. These tests were originally developed by projects under the National Defense Research Committee.

Follow-up studies of the early classes in tactical radar training disclosed that the Ratio Estimation and Coordinate Reading Tests were of little value in predicting success; the Polar-Grid Coordinate and Relative Movement Tests gave promise of being useful if revised to eliminate certain defects, notably that too much time was allowed to complete the tests and that the Relative Movement Test was too short to provide a reliable score. Since a Scale Reading Test, developed by a National Defense Research Committee project had been found effective in predicting the success of radar operators (enlisted personnel charged with the operation of radar gear) it was decided to build a Combat Information Center Aptitude Test consisting of three parts: Polar-Grid Coordinate, Scale Reading, and Relative Movement. The test is of the multiple-choice type and uses a machine-scorable separate answer sheet. A description of the test follows:

	Number of Items	Time in Minutes
Test 1. Polar-Grid Coordinate		
Items in this test require that the examinee read a point on a polar chart and then translate this reading to a point on a grid chart	45	13
Test 2. Scale Reading		
The test consists of a wide variety of scales with an arrow pointing to a certain point on the scale. The examinee must note carefully the units into which the scale is divided and make a best estimate of the point on the scale at which the arrow is placed	60	12
Test 3. Relative Movement Test		
This test contains items designed to measure the ability to visualize the relative movement of ships, involving the determination of direction, distance or speed of ships. Basically the test appears to be a measure of spatial relations ability but presents problems in a navigational setting	45	25
Total	150	50

STANDARDIZATION. The CIC Aptitude Test was standardized by administering it to a population of 860 midshipmen. The average raw score in the test made by this group is 75.96 and the standard deviation is 15.05. When administered at the tactical radar training school to a group of 307 students, a mean raw score of 95.43 and standard deviation of 20.27 were obtained. In other words, the average student in tactical radar training is more than one sigma above the mean of midshipmen.

INTERCORRELATIONS. The intercorrelations of the subtests and the correlation coefficients of each part with the whole, computed from the scores made by 118 students enrolled in a single class in tactical radar training, are presented in Table 1-VIII. The data indicate that while there is substantial relationship among the parts of the CIC Aptitude Test, they appear to measure somewhat different facets of the total ability measured by the test.

RELIABILITY. Kuder-Richardson reliability coefficients, determined for both tactical radar and midshipman populations are shown in Table 2-VIII. Contrary to what one might expect, the tactical radar student population has a larger standard deviation than the midshipmen on the parts of the test as well as on the total. A much larger number of high scores was made by the tactical radar students, but

a few very low scores were also made. Since the tactical radar school received some unscreened students from ships and shore stations in addition to those selected from indoctrination and reserve midshipmen's schools, this wide range of talent is not surprising.

TABLE 1-VIII. Part-whole and interpart correlation coefficients for the CIC Aptitude Test. (N = 118)

Test	Test Part		
	Polar-Grid Coordinate	Scale Reading	Relative Movement
Part			
Scale Reading	.69		
Relative Movement	.55	.53	
Total	.87	.88	.78

VALIDITY. Complete data on the relationship between CIC Aptitude Test scores and success in training of tactical radar and fighter director officers are presented in Chapter XII. Briefly, for tactical radar officers the correlation coefficients ranged from .45 to .56 and for fighter directors (Naval Radar Training School) from .42 to .55. Modifications of the three parts of the test were administered to a

TABLE 2-VIII. Reliability coefficients[1] for the parts of and the total CIC Aptitude Test

Test	Sample	
	Reserve Midshipmen (N = 860)	Tactical Radar Student Officers (N = 307)
Part		
Polar-Grid Coordinate	.67	.85
Scale Reading	.74	.85
Relative Movement	.70	.82
Total	.84	.92

[1] Kuder-Richardson formula, No. 21.

class of 126 radar operators (enlisted personnel). Correlation coefficients for this population between scores on Parts 1, 2, and 3 of the test and final grades were .40, .46, and .24, respectively. The Relative Movement Test in modified form was found by a project of the National Defense Research Committee to correlate .41 with success in training of sonar officers.

EVALUATION. The CIC Aptitude Test has demonstrated its value in predicting success of officers in training for tactical radar and fighter director billets. It also appears likely that the test would be effective in predicting success in training of sonar officers. Tactical radar, fighter director, and sonar officer billets are operational rather than technical in character, requiring quickness of judgment and the ability to visualize a tactical situation. For the purpose of identifying those individuals most likely to succeed in these training programs, the CIC Aptitude Test would seem to possess considerable merit.

Pre-Radar Officer Aptitude Test Battery

The supply of officers with a civilian electronics engineering background of a caliber to qualify them for direct commissioning as radar specialists was exhausted in the early months of the war. Thereafter, it became necessary to select from general officer candidate material those men who, in terms of ability, educational background, and interest, gave most promise of succeeding in the training program for radar officers. For this purpose a battery consisting of a special aptitude test and achievement examinations in general physics and general mathematics was constructed for administration to graduates of midshipman training.

An experimental form of the battery was first administered to 3,411 midshipmen at three reserve midshipmen's schools. From this population a representative sample of 400 test papers was selected for statistical analysis. In general, items which were retained in the final battery discriminated between individuals whose total test scores were in the upper 27 per cent of the sample and individuals with total scores in the lower 27 per cent.

The tests of the Pre-Radar Officer Aptitude Test Battery as finally developed are as follows:

Name of Test	Number of Items	Time in Minutes
Pre-Radar General Aptitude Test		
Part 1, Mechanical Aptitude	57	30
Part 2, Number Series	30	15
Part 3, Analogies	50	30
Total	137	75
General Physics Test	70	90
General Mathematics Test	45	75

STANDARDIZATION. The tests comprising the pre-radar battery were not standardized in the conventional sense. The chief purpose in giving the tests was to identify the upper end of the distribution of scientific ability. Final selection was based upon interview data and a careful evaluation of the candidate's educational and vocational background. A cutting score in raw score form was found to be satisfactory in this selection process.

RELIABILITY. Kuder-Richardson coefficients of reliability for the Pre-Radar General Aptitude Test, General Physics Test, and General Mathematics Test and for the total battery are .83, .76, .67, and .87, respectively. These coefficients were based upon the scores made by 3,007 general duty midshipmen in training at three reserve midshipmen's schools, the type of population to which the tests were administered in the selection of officers for pre-radar training.

TABLE 3-VIII. Means and standard deviations for the Pre-Radar Officer Aptitude Test Battery administered to general duty and specialist duty midshipmen

Midshipman Group	N	Test: Pre-Radar General Aptitude M	Pre-Radar General Aptitude σ	General Physics M	General Physics σ	General Mathematics M	General Mathematics σ	Total Battery M	Total Battery σ
Specialist	404	94.7	15.7	44.2	10.6	24.8	7.4	163.6	28.6
General	3,007	87.4	13.3	28.0	8.2	18.3	5.6	134.9	22.0

VALIDITY. Validity studies of the Pre-Radar Officer Aptitude Test Battery are of two general types:

1. Studies showing differences in average test scores of scientifically and non-scientifically trained individuals, and
2. Studies showing the relationship between the test scores and success in pre-radar training.

The first type of study is illustrated by the results obtained by administering the test battery to 3,411 midshipmen at three reserve midshipmen's schools. Of the total sample there were 404 men who were classified as specialists by virtue of their having had previous studies of a scientific nature. The remaining 3,007 men were classified as "general" and were being trained for general deck duties. Comparison of the mean scores of these two groups on the Pre-Radar Officer Aptitude Test Battery as shown in Table 3-VIII, shows a distinct superiority favoring the specialist group and, in a sense,

indicates the validity of the examinations. All differences are in favor of the specialist group and are well above the one per cent level of statistical significance.

The second type of validity evidence is illustrated by the study made of the relationship between Pre-Radar Officer Aptitude Test Battery scores and success in pre-radar training as measured by final grades. The correlation coefficients obtained in this study, for 120 cases, are as follows:

Test	r
Pre-Radar General Aptitude	.09
General Physics	.54
General Mathematics	.48
General Physics and General Mathematics	.59
Pre-Radar Officer Aptitude Battery (Total)	.45

From these data it will be noted that the Pre-Radar General Aptitude Test was the least efficient test of the battery. In fact, when combined in a multiple score, its erratic predictions had the effect of lowering the correlation coefficients which were obtained by the use of either one of the other two tests by itself. As might be expected, the General Physics Test was the best single predictor, since so much of the training course is based upon this subject matter. Actually, the 120 cases in the above sample represent a restricted range of talent since original selection took place at reserve midshipmen's schools and further personnel attrition occurred during the four months pre-radar course.

EVALUATION. The General Physics and General Mathematics Tests of the Pre-Radar Officer Aptitude Test Battery have proved to be effective instruments in the selection of officers for pre-radar training. The fact that the Pre-Radar General Aptitude Test shows a low correlation with success in training of pre-radar officers is indicative of the ineffectiveness of "general" tests in the selection of candidates for this type of training. The General Physics and General Mathematics Tests, which are in reality achievement examinations, are much more effective as selection devices at this level of ability.

The Radio Technician Selection Test

The training program for radio technicians has been one of the longest and most rigorous of the naval programs for enlisted personnel. In the early months of the war only men who had some previous experience in radio and electricity were accepted for training.

To insure the maintenance of personnel quality the Radio Technician Selection Test, in large part a measure of achievement in radio materiel, was developed. As the supply of men experienced in radio diminished, and it became necessary to select for training from the general population, the early forms of the Radio Technician Selection Test decreased in validity. The present series, Forms 6A, 7A, 8A, 9A, was consequently developed to meet the very real need for identifying those individuals who, although lacking in practical experience, possessed not only the necessary mental ability, but also enough of a previous background and interest in scientific subject matter to have become familiar with many of the elementary electrical and radio concepts. Each of the presently used forms of the Radio Technician Selection Test has the following content: Mathematics —20 items; General Science—20 items; Shop Practice—10 items; Electricity—15 items; Radio—15 items. Each item in the Mathematics section is given a weight of two points; the other items count one point, for a total score of 100 points. The test has been administered both in recruiting stations and in naval training centers, with a single time limit of 75 minutes.

STANDARDIZATION. Since the purpose of the Radio Technician Selection Test has been to identify those men in the upper 25 to 30 per cent of the enlisted population who could master the radio materiel course, the chief object of standardization of each form has been to establish a suitable cutting score rather than to establish norms of performance for enlisted men generally. This objective has been accomplished by administering the new form and an old form (previously standardized) to a population of 500 or more men who were candidates for radio materiel training. These men were selected on the basis of their scores (Navy Standard Score of 55 or above) in the General Classification Test, Arithmetical Reasoning Test, and Electrical Knowledge score of the Mechanical Knowledge Test of the Basic Test Battery. In order to neutralize practice effects, half of the men took the old form first, followed by the new form, and the other half of the men took the tests in reverse order.

The distributions of scores for the new form and old form administered to the same population made possible the equating of raw scores and the establishment of a comparable cutting score for the new form. Three ranges of cutting scores were established for each form—a passing range, a failing range and an in-between or alternate range. Records of men scoring in the alternate range were evaluated to determine whether the individuals possessed contributory background factors which might insure success in training even though aptitude test scores alone did not indicate a high level of potential achievement. All forms now in use have been equated and

comparability carefully established so they may be used interchangeably.

Reliability and Comparability of Forms of the Test. Kuder-Richardson reliability coefficients of the various forms, based on representative samples of naval enlisted personnel, ranged from .90 to .95. Coefficients of correlation between the various forms administered under conditions equalizing practice effects ranged from .83 to .89.

Correlation with Basic Test Battery. In general it has not been feasible to study the interrelationship of the Radio Technician Selection Test with other Navy selection tests because trainee samples in the radio technician training program were so highly selected that the restricted ranges would result in correlation coefficients spuriously low. Coefficients of correlation between scores on the Radio Technician Selection Test and scores on the basic battery

Table 4-viii. Correlation coefficients between scores on the Basic Test Battery, Form 3, and the Radio Technician Selection Test, Form 9A, for an unselected sample of Regular Navy recruits. (N = 534)

Test	*r*
General Classification	.55
Reading	.54
Arithmetical Reasoning	.59
Mechanical Aptitude	.47
Mechanical Knowledge (Mechanical Score)	.37
Mechanical Knowledge (Electrical Score)	.54

tests were obtained, however, for an unselected sample of recruits who entered the regular Navy after the surrender of Japan. These correlation coefficients (Table 4-viii) may be considered representative in view of the fact that the ability range of the recruit sample more closely approximates the normal distribution than does that of the selected radio technician population. It should be noted that for this sample the mean scores on tests of the Basic Test Battery run slightly lower (Table 5-viii) than those obtained for the Naval Reserve populations enlisted through Selective Service during the war. The standard deviations for the sample are also slightly smaller.

Validity. The best evidence of the validity of the Radio Technician Selection Test is seen in the fact that its use in selection has aided in the reduction of personnel attrition in radio materiel schools. Any effort to make a conventional correlational validity study of these tests is beset with the difficulties caused by restricted variability of samples. The original selection of radio technician

trainees removed the lower 75 to 80 per cent of the total recruit distribution, and of the remaining group, approximately 25 per cent were lost during the first month of instruction (Pre-Radio Materiel School). In the succeeding three months (Elementary Electricity and Radio Materiel School) another 7 to 10 per cent were lost. Correlation coefficients must, therefore, be interpreted in the light of this restricted range. For one sample of 437 men who were tested with both Forms 6A and 7A of the Radio Technician Selection Test, the product-moment correlation coefficients with Form 3 of the Pre-Radio Materiel Achievement Examination were .31 and .57, respectively. The correlation coefficients between scores on Forms 6A and 7A (for a group of 165 men) and scores on Form 4 of the Elementary Electricity and Radio Materiel Final Achievement Examination were .34 and .44, respectively.

TABLE 5-VIII. Means and standard deviations for the Basic Test Battery, Form 3, for Navy inductees and Regular Navy recruits

Test	Inductees Tested February through August 1945 (N = 260,000)		Regular Navy Recruits Sample Tested October 1945 (N = 534)	
	Mean	σ	Mean	σ
General Classification	51.99	11.85	49.51	9.70
Reading	53.51	12.53	50.93	10.93
Arithmetical Reasoning	51.62	12.40	48.01	10.44
Mechanical Aptitude	52.77	9.26	52.47	8.92
Mechanical Knowledge (Mechanical Score)	49.80	9.86	47.44	8.76
Mechanical Knowledge (Electrical Score)	51.86	10.59	48.54	8.75

EVALUATION. The Radio Technician Selection Test has been of marked value in identifying those individuals in the upper levels of ability of enlisted men who by interest and aptitude were good risks for radio materiel training. Since radio technicians were sorely needed by the Navy it was necessary to select men who could master a difficult course in a comparatively brief span of time. The fact that attrition was relatively small in the radio materiel training program can be attributed in part to the fact that the trainees were rigorously screened by means of this test.

The Radio Code Test

The Radio Code Test—Speed of Response was developed for the Army and Navy by the National Defense Research Committee. The need for some type of special aptitude test was indicated by the high attrition rate in radio schools and the fact that tests of intellectual

capacities consistently showed low correlations with success in training in such schools. Since the reason for failure most frequently mentioned by the staffs of radio schools was "inability to learn the code" it appeared that, if possible, a special aptitude test should be designed to predict code learning ability.

In developing the code aptitude test for the armed services, work was undertaken on (1) *discrimination* tests measuring the ability of the subjects to distinguish between complex rhythmic patterns of dots and dashes, (2) *code learning* tests which measure the ability of the subject to learn code characters, and (3) *speed of response* tests which measure the ability of a person to identify a few code characters at rapid rates of transmission, the subject being first taught a few code characters and being then tested on these well-learned characters at several faster code speeds. Theoretical considerations and exploratory studies indicated that the speed of response type of test showed the greatest promise of being useful in predicting success in radio training and a test embodying this principle was, therefore, developed. A complete account of this project has been presented in reports of the National Defense Research Committee.

The form of the test adopted for use by the Navy consists of two parts, first, a learning unit in which the subjects are taught three code characters, and second, a testing unit in which the subjects are tested at four different speeds on these learned characters. The whole test is recorded on phonograph records and the examinee's responses are recorded on machine-scorable answer sheets. Only the two highest speeds, consisting of 75 items each, are actually scored.

STANDARDIZATION. The Radio Code Test—Speed of Response was standardized on a Navy recruit population and the raw scores converted to Navy Standard Scores (mean of 50 and standard deviation 10). The test is regularly administered to recruits at naval training centers as a part of the Basic Test Battery.

VALIDITY. Extensive validity data on the test have been published by the National Defense Research Committee. Correlation data gathered routinely by the Bureau of Naval Personnel are presented in Chapter XII. In a special study the following correlation coefficients between scores on the Radio Code Test—Speed of Response and various criteria were obtained: .37 with the fourteenth week code speed grade; .28 with final course grade; .49 (biserial r) with a pass-fail criterion. The General Classification Test was found to correlate .12, .15 and .09 with the three criteria, respectively. If a cutting score of 55 (Navy Standard Score) on the Radio Code Test had been applied to the class studied, two-thirds of the failing students would have been eliminated and attrition would have been reduced from 30 per cent to about 17 per cent.

EVALUATION. The Radio Code Test—Speed of Response has demonstrated its value in the selection of candidates for radio school. Through its use it has been possible to identify more clearly those individuals who, while they might be only average in general mental ability, possessed special ability to learn the code. While the test did not reduce failure rate in radio schools to the low percentage found in most other enlisted elementary schools, it did constitute a distinct improvement over the original tests of the Basic Test Battery as a selection device.

Group II Tests

THE SONAR PITCH MEMORY TEST was developed for the Navy under the direction of the National Defense Research Committee. The test is basically similar to the Seashore Pitch Test but contains certain improvements which make the test more similar to the actual job of a sonar operator and which also increase its reliability. The test was standardized on a population of Navy trainees all of whom had General Classification Test scores above 50 (Navy Standard Score). Substantial correlation coefficients with the doppler judgment criterion have been obtained. The test has been used as part of a selection battery to select sonar officers and sonar operators. In the case of officers, the test has been given at reserve midshipmen's or indoctrination schools or in operational training centers; enlisted men take the test at naval training centers as a part of the general selection process. Complete reports describing the development of this test and its validity have been published by the National Defense Research Committee.

THE WINCHMEN AND HATCHMEN SELECTION TEST was developed for use in the selection of winchmen and hatchmen for the training program under the direction of the Operational Training Command, Pacific Fleet. The test consists of three parts as follows: Test 1, Block and Pattern Assembly, 34 items; Test 2, Rotation of Solid Figures, 15 items; Test 3, Mechanical Comprehension, 40 items. The time limits for the three parts are 15, 12, and 20 minutes, respectively. The total score in the test is the number of items answered correctly. Correlation coefficients with criteria of success in operational training have been uniformly low; and when used in connection with the best tests of the Basic Test Battery, there is no resulting increase in efficiency of prediction.

THE EYE-HAND COORDINATION TEST was developed also for use in the winchmen and hatchmen selection program. The test is a measure of speed and accuracy of serial reaction and consists of a series of circles, connected by lines, which the examinee marks in serial order. There are 120 circles in the test and a time limit of

60 seconds is allowed. A study of the use of the test in the selection of winchmen and hatchmen showed that the multiple correlation coefficient obtained by use of the best combination of two tests from the Basic Test Battery was not improved by the addition of this test.

THE AIRPLANE MATCHING TEST was prepared from material originally developed by the Aviation Psychology Section, Bureau of Medicine and Surgery, U. S. Navy. The test consists of plane silhouettes, partially covered by cloud formations, which the examinee is required to match with the contour silhouettes of 14 types of planes shown on a master chart. Eight minutes are allowed to complete the test of 40 items. The score is the number of rights minus wrongs.

The test has been used at the recruiting level in the selection of combat aircrewmen as a supplement to the Basic Test Battery. In a study of the prediction of success in aerial gunnery school it was found that the multiple correlation coefficient, using the Mechanical Aptitude, Mechanical Knowledge, and Reading Tests of the basic battery was .38; the addition of the Airplane Matching Test and the Mechanical Comprehension Test (described below) raised this coefficient to .41. Considering the present degree of refinement in the selection process, it does not appear that the addition of the Airplane Matching Test and Mechanical Comprehension Test to the Basic Test Battery results in an appreciable increase in efficiency of prediction of success, at least as measured by grades in aerial gunnery school.

THE MECHANICAL COMPREHENSION TEST used in the selection of combat aircrewmen is the Bennett Mechanical Comprehension Test. Since this test is adequately described in the literature, no description of it is presented here.

THE DIGIT MEMORY SPAN AND SENTENCES IN NOISE TESTS, developed by the National Defense Research Committee, are of the recorded disc type and are used in the selection of telephone talkers at some of the advanced classification centers. The Digit Memory Span Test consists of 24 items, extending from the four-digit to the ten-digit level. The test exists in two forms: the Full Test, on two recorded sides, consists of 24 items and can be administered in 13 minutes over a loudspeaker, or 15 minutes on earphones. The Short Test, on two recorded sides, consists of the first 18 of the 24 items in the long form and can be administered in 11 minutes over a loudspeaker, or 13 minutes over earphones. The subject records his responses in appropriate blank spaces on an answer sheet. The score in both forms of the test is the number right.

THE SENTENCES IN NOISE TEST is designed to measure the individual's ability to understand messages spoken over military communications systems in the presence of interfering noises. It is made

up of 100 items consisting of questions, commands, and incomplete statements, for each of which four alternative words or numerals are given, one of which serves to complete correctly the stimulus sentence. The listeners answer by underlining the appropriate response. The test requires about 25 minutes and can be administered over earphones or a loudspeaker. The score in the test is the number right.

No studies of the validity of the Digit Memory Span or Sentences in Noise Tests have been conducted by the Bureau of Personnel. Complete descriptions of and statistical reports on the tests have been published by the National Defense Research Committee.

Usefulness of Special Aptitude Tests

On the basis of data gathered in the Bureau of Naval Personnel during World War II, it would be difficult to defend the extensive use of special aptitude tests of the pencil and paper type in the selection and classification of naval personnel. Such basic classification batteries as the Basic Test Battery for enlisted personnel and the Officer Classification Battery appear to be adequate for the testing purposes to be served within the framework of the present rather unrefined selection and classification programs. The major exceptions occur in the testing of personnel for selection for radio and sonar training, where tests of intellectual and mechanical aptitudes are inadequate. Special tests have also been useful in selecting officers and enlisted men at the upper end of the distribution of technical ability for such training programs as pre-radar and radio technician. In brief, special tests serve a useful purpose when (1) the measurement of highly specialized aptitudes is required, and (2) when refined measurement is required at a given level of ability.

It should be emphasized that the special aptitude tests studied by the Bureau of Naval Personnel have been predominantly of the paper and pencil type. It seems likely that for certain billets, notably in the artificer branch on the enlisted level, the use of performance tests might have resulted in greater efficiency of prediction. As selection and classification procedures as a whole are refined, it would appear that greater refinement in the aptitude testing area should be sought by determining the usefulness of special aptitude tests of the performance type for various billets especially in the artificer branch.

CHAPTER IX

MEASURES OF PERSONAL ADJUSTMENT

A MILITARY organization cannot afford to harbor the maladjusted individual. In a time of crisis his disintegration is apt to produce disastrous panic by contagion. His failure of judgment or courage may prove the failure of a ship or an operation. His emotional instability or lack of initiative may destroy the balance and coordination which is the essence of a complex interdependent fighting force. Unless every man can be counted on to do his job under the stress of battle, the efficiency of the whole fighting team will be impaired.

Unfortunately the problem of eliminating psychoneurotic and psychotic elements from among the populations drawn into the armed services has proved both difficult and extensive. Analysis of the data gathered at the level of the local board and the induction center indicates that an average of 28 men per 1,000 have been rejected because of mental disease, not including 7 men per 1,000 disqualified because of mental deficiency. Further elimination has occurred at training centers. It has been estimated that administrative disposition for from two to three per cent of all Navy recruits has been necessary on psychiatric grounds.[1] The extent of the problem may be judged by the provision contained in a joint letter from the Chiefs of the Bureau of Medicine and Surgery and Bureau of Naval Personnel directing Senior Medical Officers to provide 35 beds per 1,000 incoming recruits as a psychiatric observation ward except where experience indicates a lesser need.

The breadth of the problem has been matched by the difficulty of its solution. With tremendous numbers of men entering the naval service, the shortage of trained psychiatric and psychological personnel became acute. This lack was further complicated by the inevitable enlargement of the areas within which psychiatric evaluation and screening was necessary. Selection for the submarine service, as a typical instance, called for special clinical appraisal and cast an additional burden upon psychiatric personnel. As the war progressed, the problems of combat and operational fatigue became more numerous, and psychiatric interviewing at redistribution centers and receiving ships assumed increasing significance.

The urgent need for test procedures arose primarily out of the existence of the difficult situation in which a relatively few expe-

[1] Since the measures of personal adjustment described in this chapter were used as "psychiatric screening devices," this term will frequently be applied to them in this discussion.

rienced clinicians were required to undertake the task of diagnosing and making recommendations with respect to millions of men and women. The goal was set as the development of objective test techniques suitable for group administration which would operate as an effective adjunct to the psychiatric screening process.

USES AND MISUSES OF THE SCREENING TESTS. The primary function of the screening tests developed in the Navy has been to select that segment of the total population in need of referral for psychiatric investigation. The tests must therefore be considered adjunctive instruments and do not usurp the functions of the psychiatrist. Only under very special circumstances have measures of personal adjustment been permitted to operate as a direct means of elimination in the screening process without confirmation through independent clinical interviewing. In line with this thinking there has been a general tendency to eliminate the use of the term "cutting score" in connection with these tests and to substitute the terms "referral score" or "investigating score." In some instances the tests have not even been used to restrict the population to be interviewed but made to serve primarily as an indication of the individuals to be examined most intensively. Under either circumstance, it should not be assumed that interviews could proceed on the leisurely basis which characterized civilian psychiatry. The five-minute interview, at least for training centers and receiving ships, frequently represented an extravagant indulgence of time on a single case. Techniques for group interviews, in which a number of men were seen at one time, have been developed in several places. With or without the aid of pre-screening tests, rapid if not cursory evaluation remained a prime necessity.

This situation points to a second important function of the paper and pencil questionnaire used in measuring personal adjustment. While it is far from universally true, a great many interviewers have found that the data provided by such a test assisted materially in condensing the examination and pointing it in a productive direction. Not only were the specific admissions concerning symptomatology significant, but the subject's variable reaction to the relatively impersonal test and to the face-to-face questioning of the psychiatrist often provided important clues to areas of disturbance.

This use of the test data as a basis for interviewing does not always constitute an unmixed blessing. In the rush of events, when time was short and waiting lines of men long, the test might easily slant the interview and even determine its whole nature. The examination might readily assume a rigid pattern of questioning wholly fixed by the symptomatology covered by the test, or the test scores could be overvalued in making a final disposition of a case. These disad-

vantages, however, are functions of the skill and flexibility of the individual examiner. Stereotyped and formalized approaches to a subject could also occur without the predisposing influence of test data.

HISTORY OF RESEARCH IN THE BUREAU OF NAVAL PERSONNEL ON PERSONAL ADJUSTMENT MEASURES. The mental health and stability of naval personnel is the practical concern of both medicine and personnel administration. For this reason, the Neuropsychiatric Branch of the Bureau of Medicine and Surgery and the Test and Research Section of the Bureau of Naval Personnel have maintained active liaison and engaged in a joint endeavor to reach a satisfactory solution of the problem of developing effective screening tests. While the psychologists assigned to the Bureau of Naval Personnel have assumed primary responsibility for actual research in test development and validation, cooperation by the psychiatrists and psychologists assigned to the Bureau of Medicine and Surgery has been of the utmost value.

The measure of the joint interest is seen in the fact that one of the tests in general use in the Navy, the Personal Inventory, was installed for use through the joint recommendation of the Chiefs of the two Bureaus. Numerous conferences involving representatives of the two Bureaus have marked the course of the research on test development and validation. Of special importance to the research program have been the contributions made by psychiatric personnel in providing criterion measures for the validation of tests. Whatever progress has been made in the development of effective screening tests stems directly from the coordinated interests and efforts of the representatives of the Bureau of Medicine and Surgery and the Bureau of Naval Personnel.

It should not be imagined from what has been said that the plan of research on screening tests always represented a coordinated large-scale attack on the problem. On the contrary, the program was frequently of an adventitious nature. To evade recognition of this bit of history or to deny its implications would be to do a disservice to the future.

There has never been in the Navy an organized, thoroughly planned, adequately staffed program designed to exhaust the possibilities of available methods for the identification of maladjusted individuals by group test procedures. In terms of the extent and seriousness of the problem, the research which was undertaken during World War II can unfortunately be characterized by the overworked phrase "too little and too late." An explanation for this situation is reasonably clear. So many problems of an urgent nature existed that compromises with quality and quantity were necessary.

Beyond that, the problem of personality evaluation is still one of the most perplexing on the psychological research agenda. It is not surprising, therefore, that other pressing questions received prior attention.

These comments are injected here because the armed services, even after the termination of hostilities, offer a laboratory with unequaled opportunities for the study of psychological disturbances and for the development of techniques for their detection. It is a matter of practical rather than academic interest which should prompt the military to investigate these problems on a large and intensive scale. Barring the eventuality of "thirty minute wars," one of the prime personnel problems for a military organization to solve will still be the adequate evaluation of the adjustment of both its officer and enlisted personnel.

The earliest organized effort in the Navy to secure a test to aid in the psychiatric screening of personnel came with the establishment of a National Defense Research Committee project. It was not until June 1943 that a project report was published presenting details of the test developed under this project, a questionnaire entitled the Personal Inventory. Subsequent research on this instrument produced a variety of significant changes, including a reduction in the number of items from 145 to 20. For all practical purposes an effective and usable instrument with adequate validity data was not available for general use by the Bureau of Naval Personnel until February 1944. This statement, of course, does not take into account locally initiated test development or validation projects designed to assist in the screening of psychiatric cases.

The role played by the Test and Research Section of the Bureau of Naval Personnel came into being much later. Although the psychologists assigned to this Section had been actively interested in the progress of the work of the National Defense Research Committee project and had been consulted in this connection on numerous occasions, they did not begin direct participation in research on psychiatric screening instruments until September 1944.

Activation of the research program within the Bureau came initially by way of a request from the Enlisted Classification Section for a test which could be used in screening WAVES applying for overseas duty. This request was followed almost immediately by requests for further research on screening instruments in connection with assignment of personnel to submarine and amphibious training. With the development of new tests and the revision of established inventories, further data were required with respect to the general recruit population. Projects were then set up to cover this broad population in the training centers. The problem of psychiatric

screening of reserve midshipmen was raised by the Officer Selection Unit and a brief study of this question completed. Specific investigation of test procedures for the screening of overseas veterans exhibiting signs of combat or operational fatigue was begun after a need for such an instrument was indicated by correspondence from the commanding officer of a West Coast receiving ship, and from the Chief of the Bureau of Medicine and Surgery.

Personnel from the Test and Research Section were assigned to the various projects as the pressure of work increased. Since the development of the various research projects depended largely on the receipt of specific requests in connection with particular problems arising in other sections of the Bureau, or at various training centers or receiving ships, no overall research program was developed. This will account, in part at least, for a certain discontinuity in the research undertaken and for the neglect of certain populations and problems which are undoubtedly significant for the general issues investigated.

The instruments described in this section do not represent all the psychiatric screening instruments developed either within the Navy or by such contractual agencies as the National Defense Research Committee. Only those tests used in connection with research projects undertaken by the Test and Research Section of the Bureau of Naval Personnel are described. While research data will be presented on only four of the instruments described below, descriptions of the others are included in order to indicate the different types of instruments developed and studied.

Enlisted Personal Inventory, Form 2. This inventory contains two parts, Part 1 being a condensed form of the 145-item Enlisted Personal Inventory, Form 1, developed by the National Defense Research Committee. Part 2 is the same as Form N of the Cornell Selectee Index.

More specifically, Part 1 contains 20 items, most of which are cast in forced-choice format, such as,

I wish I weren't bothered by bad dreams.	____	____ I wish I could have more excitement.

The term forced-choice is applied because the alternative statements, which are presented on opposite sides of the answer sheet, do not lie on any psychological continuum. The subject is required to choose between conditions which are not really alternatives, both of which may conceivably be inapplicable to his own experience. In connection with the illustrative item above, it is entirely possible that the subject never dreams and has all the excitement he cares to have. Some of the items of this test are in paired-choice rather

than forced-choice form and do present statements which represent a continuum in the sense that they deal with different intensities of the same psychological symptom. For example,

When excited I feel weak. ____ ____ When excited I feel stronger.

Still other items are cast in yes-no form, which, in reality, is only a special form of the paired-choice item.

Part 2 (Cornell Selectee Index) consists of 32 items, all in yes-no format. Since Parts 1 and 2 were developed independently, there is some degree of overlap in the content of the items but this is not excessive. It was originally intended to score only 16 of the 32 items included in Part 2. These had been designated by the authors of the instrument as stop items. The function of the stop item, as originally described, was to make each single scored item crucial for screening. If any single one of the symptoms described was affirmatively reported by the subject, he was to be eliminated or referred for interview, depending on how the test was to be used. Quite naturally, these items tend to deal with symptomatology of great severity and significance from the diagnostic point of view. It is difficult to pick a representative item but the following will serve to indicate at least the typical form.

Have you ever had a fit or a convulsion? Yes No

The entire test is printed on an IBM answer sheet and can be scored by either machine or manual means using a scoring stencil.

Officer Personal Inventory, Form 1. This 164-item questionnaire was developed by the National Defense Research Committee and contains material similar to that found in Part 1 of the Enlisted Personal Inventory, Form 2. In general the forced-choice item format prevails. It is highly probable that future research will indicate the wisdom of reducing the length of this test to something comparable to that in the test used with enlisted personnel. It has been a repeated finding of most investigators that approximately the same results can be obtained with relatively few items and that extreme length in such an inventory adds little to its validity.

Item analysis revealed that 50 items discriminated between maladjusted and normal groups, and a scoring key has been developed using these items. The test is presented in a printed booklet and separate IBM answer sheets for machine or hand scoring are available.

Personal Inventory, Form X-1(W). This test, which is substantially the same as the Officer Personal Inventory, Form 1, was prepared for use with enlisted WAVES. Only those questions in the

original test which were specifically inapplicable to a female population were altered. Whenever possible, reasonably comparable questions were substituted for those requiring change. When the content was altogether inapplicable to a female group, a substitute question was chosen from the Enlisted Personal Inventory.

The officer rather than the enlisted form of the Personal Inventory was adopted for WAVES use because the former is written in more sophisticated language. Since the educational level of the enlisted WAVES group is somewhat higher than that of male recruits it was believed that an adaptation of the officer form would prove more satisfactory.

BILLET QUALIFICATIONS BLANK, FORM X-2(M). This experimental inventory was developed for use with enlisted men in the general recruit population as well as for those being assigned to submarine and amphibious duty. To understand its basic format and content it is necessary to appreciate that malingering in the sense of the exaggeration of symptoms is a reasonably rare phenomenon in the military services. When it does occur, psychiatric scrutiny seems indicated at all events and no loss is incurred by referral of the subject to the psychiatric unit. On the other hand, a careless attitude toward the test or a positive effort to produce an exaggerated picture of satisfactory adjustment are much more common. It is in relationship to such attitudes that the value of the previously discussed forced-choice type item must be assessed. In the first place, it permits alternation of answers so that the choice indicating good adjustment may appear on the left or right side of the page. This compels more attention and prevents a casual marking "down the line" which is reasonably frequent in the yes-no format. In addition, since the alternatives are discontinuous in the sense of a cohesive or related psychological content, the person tested may actually have some difficulty in choosing the answer least indicative of maladjustment. This may be clearly seen in such an item as one which requires the individual to indicate whether he has suffered more from head colds or more from dizziness. Without regard to the resistance or hostility such an item may engender in the subject, it plainly makes difficult a perseverative indifferent test attitude.

With a view to securing more careful introspection and more accurate reporting, and also to permit some correction for whatever exaggeration of good adjustment did exist, three features were incorporated in the Billet Qualifications Blank.

1. Many of the items were placed in a billet context on the theory that most personnel in the Navy, especially recruits, have some anxiety with respect to their assignments. Instead of asking whether the subject is a person who gets nervous or dizzy when working in high places, he

is asked whether he thinks he is qualified or disqualified for a billet in which freedom from dizziness in high places is essential. In this way the anxiety associated with impending billet assignment is used as a pressure to produce more accurate and careful reporting. Both in the directions for administration and in other incidental features connected with the framing of the questions, the notion of the significance of the test for assignment purposes is kept constantly before the subject.

2. A "self-idealization" scale has been included for the purpose of assessing the degree to which the individual attempts to exaggerate good adjustment. This scale is typical of the "lie" scale which may be found in other psychological tests of personal adjustment. In this test, however, the items are cast in the billet context. For example, a typical item calls on the subject to qualify or disqualify himself for a billet requiring an individual who *never* experienced either fear or anxiety. It was hypothesized that a score on such a scale might act successfully as a suppressor variable when combined with other scores derived from the part of the test measuring maladjustment directly. If the subject acted defensively toward the test and tended to exaggerate or idealize his adjustment, it was supposed he would receive a high score on the "self-idealization" scale which would operate in a measure to correct the distorted low score on the scale measuring maladjustment. There are many complex, theoretical problems involved in the concept of a suppressor variable to correct scores on personality tests. These cannot be discussed in detail here, but the general possibility of raising the multiple correlation between a battery of tests or parts of a test and a criterion by including a test or subtest with zero correlation with the criterion has received previous attention in the psychological literature.[2]

3. A third feature of the test involves inclusion of items dealing with attitudes toward specific experiences both prior to and since enlistment in the Navy. It has been recognized that a purely symptom approach to the measurement of adjustment, leaning heavily on the admission of a series of psychosomatic disabilities, leaves much to be desired. On the theory that the personal beliefs and social attitudes of the neurotic and psychotic differ markedly from those of the better adjusted individual, a series of multiple choice items was included in the test. The following is an instance of this type of item:

If a man shows that he's nervous about going to sea

(a) he probably feels like a lot of other fellows
(b) he ought to be kicked out of the Navy
(c) he ought to be given a shore billet

The test contains 100 items and provision has been made for a separate answer sheet which may be scored by hand or machine. The test is divided into three parts as follows:

[2] Horst, Paul. "The Role of Prediction Variables which are Independent of the Criterion," in Horst, Paul (Editor), *The Prediction of Personal Adjustment*, pp. 431-436, Bulletin 48, Supplementary Study E, Social Science Research Council, New York, 1941.

Part 1. Fifty items in which the individual is to qualify or disqualify himself for a billet in terms of the described characteristics. Ten of these items compose the "self-idealization" scale to which previous reference has been made.

Part 2. Twenty-five items in which psychological and somatic symptoms are listed and the subject is required to indicate the frequency with which he experiences such symptoms. The answer sheet allows for the following responses: "Never," "At times," and "Often." Repeated instructions presented in this part indicate the importance of the answers for billet assignment.

Part 3. Twenty-five items dealing with attitudes toward specific experiences both prior to and since enlistment in the Navy.

BILLET QUALIFICATIONS BLANK, FORM X-2(W). There are no substantial differences between this form, developed for use with enlisted WAVES applying for overseas duty, and Form X-2(M) described in the preceding section. The differences that do exist relate solely to those items which can have specific application only to one sex.

EXPERIENCE COMPARISON INDEX, FORM X-1. This 63-item test contains many of the customary questions dealing with psychological and somatic symptoms. A somewhat greater emphasis than usual is placed on those symptoms which, for general clinical reports, have been presented as associated with combat fatigue. This emphasis is due to the fact that the test was prepared in response to a specific request for a screening instrument which would be effective with men returning from overseas duty.

The principal feature which differentiates this test from the other inventories is found in the method of item presentation. It was supposed that men would more readily admit to neurotic symptoms if the question or item was so worded as to imply that such a condition was common to a group rather than individual and unique. Therefore men were not asked directly whether they suffered from some particular symptom such as being sensitive to noises. Instead they were given preliminary instructions to the effect that various groups of men had reported many different experiences and feelings and that the purpose of the questionnaire was to permit the subject to compare his own experiences and feelings with those of other men. A typical item would then read:

Group 6 men are very sensitive to noises. Yes No

The subject would answer "yes" if his own feelings or experiences were like those of the group described, and "no" if his experience had been different.

PERSONAL CHECK LIST, FORM X-4. This is a 57-item test in which all the items are cast in paired-choice format. Like the Experience

Comparison Index, this test was developed specifically for the purpose of screening personnel returning from overseas and therefore places heavy emphasis on those symptoms customarily associated with the syndrome designated "combat fatigue." Three separate segments may be differentiated in the test. The first 27 items have been considered as a separate unit since they were subjected to fairly critical validation by extensive research with a large population of returning veterans. Typical of this group is:

I'm restless and wakeful at night. _____ _____ I usually sleep pretty well.

A second unit is represented by 15 items similar in every way to the first 27 so far as format and reference to psychosomatic symptoms are concerned. These had not been subjected to equally rigid validation and therefore received separate treatment in the reported research.

A third unit consists of 15 experimental items somewhat similar in content to the self-idealization scale described in connection with the Billet Qualifications Blank. The form of the item has been altered to conform to the paired-choice setting of the other items in this test. Typical of this group is the following item:

I've been troubled once in a while with thoughts about sex. _____ _____ Sex ideas have never once bothered me.

BILLET PREFERENCE RECORD, FORM X-1. This 30-item test depends fundamentally on a theoretical idea first presented in the Billet Qualifications Blank but places the items in a paired choice form. A single example will indicate the general structure of the items.

I would rather have a billet–

_____(a) involving some danger and excitement.
_____(b) involving quiet routine and some monotony.

Recasting the Billet Qualifications Blank in this form proceeded on the notion that introduction of a paired-choice form would improve discrimination. Aside from the billet context of items, the other features of the earlier blank were not carried over into this later test.

PREVIOUS DUTY CHECK LIST, FORM X-1. This inventory would be better characterized as a test of attitudes toward the Navy than a measure of neurotic tendency. It was constructed on the hypothesis that hostile attitudes or griping might be associated positively with maladjustment not only toward the specific military situation but with respect to the general life situation. Section 1 contains 22 items in the following format:

Working parties were selected fairly. Yes No

Section 2 consists of five questions dealing with specific experiences such as, "Have you been injured?" No score is obtained in this part, the responses being intended for interview purposes only.

Social Judgments Test, Form X-1. This test has some of the elements of a projective technique and bears a resemblance to the procedures which characterize the multiple-choice Rorschach, although the differences are far more significant than the resemblance. To understand the technique, it is necessary to examine first the directions for administration. These indicate to the subjects that success in the Navy depends to a great extent upon the ability to size up people correctly, to read their characters, moods, and ways of acting. Along with such instructions, the subject is given a sheet on which are reproduced twenty portraits of Navy personnel, exhibiting various facial expressions. On a separate sheet is presented a series of four statements for each picture supposedly indicative of four possible moods or thoughts which might be associated with each picture. The subject is required to select the statement which he thinks best characterizes the expression on each of the portraits. The assumption is that the subject will tend to project his own moods and ideas into the situation and thus mark those statements which best represent himself.

In the construction of the items, two of the four statements given for each picture were supposedly in the direction of maladjustment. The remaining two were assumed to be within the normal or adjusted range. An example of the types of statement offered in connection with the photographs will be helpful.

_____Wish I didn't have a stomach full of butterflies.
_____The best part of this duty is the "bull sessions" with the fellows.
_____Lord, but I'm tired.
_____A guy's personal feelings are not very important in a fight this big.

Procedures in Constructing the Tests

In view of the frequent use of the phrase "measure of personal adjustment," one might assume that personal adjustment is a clearly definable concept. Actually there exist a host of diverse, even mutually exclusive, definitions. If it had been necessary to settle on the precise nature of "personal adjustment" before undertaking the measurement of its variable manifestations, we would even now be scratching fruitlessly in the boneyard of discarded phrases. Fortunately, it is not generally necessary to reach full comprehension of the thing to be assayed before attempting its assessment. A familiar illustration of this is found in the successful measurement of differences in electrical potential by means of a voltmeter despite

the fact that the precise nature of electricity is still not completely understood.

Recognition of this point is important for an understanding of the progress and failings of research in the testing of personal adjustment. If all the elements, relationships, etiology, and dynamics entering into the concept of adjustment or maladjustment were known and understood, it would have been possible to concentrate attention on the construction of tests specifically pointed at the uncovering of known and essential elements. Actually such precise knowledge does not exist. Even more, the situation possessed an additional complication. It was not general adjustment or maladjustment which was to be predicted but specific personal adaptability to military service in wartime. Though this is but a special case in connection with the total adjustment-maladjustment problem, it should be remembered that adaptability to civilian modes of existence is likewise a special case. It could not be assumed that the vastly more documented, though still inadequate, knowledge concerning personality deviations in civilian life would apply directly to the military situation.

If maladjustment cannot be sharply defined in terms of what it is, it can at least be described operationally, in terms of what it does, or in terms of the observable behavior patterns which may be reasonably said to stem from inner conflict and disturbance. Even within the specific confines of the military situation, it is feasible to express with some certainty the types of behavior which would be undesirable, if not intolerable.

This suggests that research might find its most effective point of departure in the detailed analysis of successful and failing groups within the military population to be predicted. From such analysis might come significant data concerning the background factors, attitudes, and personal or physical characteristics which would serve as predictive items in any experimental test. This procedure necessarily involves long-range planning and long-term research. It implies postponement of test development until experience with available military populations has permitted a clear differentiation of adjusted and maladjusted groups followed by a careful study of the observable and measurable characteristics of these groups.

Early in 1942 such a program might have been feasible and sound. In 1944, when the Test and Research Section began its preliminary work on psychiatric screening tests, it did not seem practical to embark on the elaborate case history studies which were indicated as a significant preliminary to test construction. In lieu of this, the development of experimental test items proceeded immediately along those lines which were indicated by immediately available evidence.

For example, some of the psychiatric literature on psychological disabilities in the military service was scanned for clues on predictive items. Similarly, the general psychiatric literature, without regard to its military content, served as a guide for item construction. Available inventories in the field of personality measurement, both in military and civilian use, were scanned for pertinent suggestions. In short, existent materials were adapted to the test hunches of the personnel assigned to the various projects, and a series of experimental forms were developed for validation.

While such a short cut procedure may have justification in a situation of wartime urgency, it cannot be supported in future research. The careful and logical prescription for an effective screening device calls for preliminary controlled studies of the characteristics of the military populations to be predicted. Instruments already developed have established their value in a very practical way. It is probable, however, that further advances in the measurement of adjustment can only come if the researcher is willing to retrace his steps and analyze the fundamental assumptions on which diagnostic and predictive items and procedures have been selected.

Research Procedures

THE CRITERION. A psychiatric screening test is designed to predict whether an individual is sufficiently well adjusted to accommodate himself to military life without significant disturbance in regard to work to be performed or in social relationships. Psychiatric interviewing is essentially designed to perform the same task. Test procedures are used primarily to save time for psychiatrists by selecting that portion of the population most in need of special interviewing.

It follows that a comparison of test results with psychiatric diagnoses is merely a comparison of one prediction with another prediction. To speak of test validity on the basis of the outcome of such a comparison is to assume, a priori, the validity of the psychiatrists' judgments. Actually agreement or lack of agreement merely serves as an index to indicate the extent to which one procedure may serve to supplement or replace another procedure. Any more profound claims for a test which agrees well with a psychiatric prediction criterion can only be supported on the basis of clear evidence that the psychiatric prediction is valid in itself.

It may well be argued that the tests, which were designed primarily as an adjunct to the psychiatric screening process, require no further validation than a measure of agreement with psychiatric diagnosis. This is true insofar as the tests are designated as measures of psychiatric prediction and insofar as no claim is necessarily made

for them as measures of personal adjustment. It is to be hoped that future research will consider the psychiatric criterion as a preliminary measure of validity and take the logical step of determining actual validity by follow-up studies of the men for whom predictions are made.

All of the research pertaining to the evaluation of the experimental instruments to be reported in this chapter has availed itself of the psychiatric criterion. While provision was generally made for further validation by case follow-ups within a period of time following test administration, the termination of hostilities and consequent demobilization has made such procedures impracticable.

It is also recognized that the mere statement that psychiatric diagnosis or rating has been made to operate as the index of test efficiency is inadequate. Whose psychiatric diagnosis? Whose rating? Given under what conditions? These are extremely pertinent questions. Psychiatrists will admit that there are significant variations in training and skill within the profession. Military necessity frequently compelled the use of unspecialized medical practitioners as neuropsychiatrists after very brief training courses. In a field in which prediction is often difficult under the most favorable circumstances, it would be manifestly unwise to accept as an accurate criterion anything less than the most skilled and experienced judgments.

To assure the soundest available criterion, research was conducted wherever possible at those stations in the Navy which were recommended by the Neuropsychiatric Branch of the Bureau of Medicine and Surgery. Where psychiatric examiners were not available, arrangements were made for the assignment of a psychiatrist on temporary additional duty. In other instances, where availability of a particular population rather than of a criterion measure dictated choice of a research station, and where special assignment of psychiatric personnel could not be made, substitute measures were taken. These included the development of external criteria through ratings supplied by classification officers, or the development of internal criteria in which a rational scoring key was first employed to segregate high scoring and low scoring groups. The significance of the differences between the high and low groups either in total test scores or for individual items was then determined. In reality this procedure made possible only an analysis for internal consistency and the validity of the items and of the test had ultimately to be determined by some adequate external criterion. In general, the substitute measures employed proved unsatisfactory.

One major difficulty in the interpretation of any particular study or in the comparison of the results from several studies relates to

what may be termed the stringency of the criterion employed. By stringency is meant the degree of maladjustment which is held necessary for inclusion of a subject in the deviated group. If the psychiatric criterion of maladjustment is set so that only the extreme deviates are included, predictive test data will almost invariably turn up with higher validities than if the criterion of maladjustment is less stringent and includes the moderately maladjusted. Since stringency in this sense can only be partly defined in advance and will depend as much on the attitudes and training of the psychiatrist as on the rating plan used, interpretation of every study must proceed with some reservations not only about the quality of the criterion but also about the level of prediction required.

RESEARCH DESIGN. While considerable variation in detail will be found in the studies which follow, the general design follows a fairly uniform pattern. In the main, tests were administered to random samples of men at a given station, to men applying for some specific assignment, or to men returning from overseas duty. These men had been interviewed or were later interviewed by a psychiatrist or a clinical psychologist with extensive experience in screening naval personnel. In some instances all men were given specific ratings on an accepted rating scale. In other instances, the regular routine of psychiatric screening was permitted to operate without special attention to those subjects tested. The adjusted and maladjusted groups were then determined by a later check of the records of the Psychiatric Unit to ascertain the individuals discharged or otherwise subjected to psychiatric disposition. Test prediction and psychiatric prediction were then correlated in various ways.

The pitfalls inherent in research in this area are more numerous and treacherous than in most psychological research. Three problems in particular, though not necessarily the most significant, deserve special mention in view of the fact that they are so generally recognized and just as generally undervalued.

Nothing is so apt to produce delusions of grandeur in the constructor of personality inventories than the use of what might be termed the "ex post facto" psychiatric criterion. This involves a technique by which a hospitalized population is contrasted with successful military personnel in a training center. The difficulty with this procedure is that patients in a neuropsychiatric ward have generally become so sensitized to clinical symptomatology that they no longer react in the way in which they would have reacted to the same instrument prior to hospitalization. Furthermore, the hospital wards generally contain a selected sample from the extreme of the deviated cases. This means an extremely stringent criterion group in the sense that the term has previously been employed. It is no remarkable thing, then that test score differentiation from a young,

enthusiastic, and healthy recruit group can be obtained with considerable ease. As a single instance, one of the tests developed in the Bureau was administered to a neuropsychiatric ward population in a naval hospital. This group was found so far different from a random sample of qualified recruits drawn from a training center that 87% of the hospital group scored above an established critical score point whereas only 4% of the qualified recruit group was found within the same range of scores. Needless to say, the test was not found so efficient under more exacting circumstances in which adjusted and maladjusted individuals were intermingled in the same group and before the deviates had been identified psychiatrically.

Comparable exaggeration of the efficiency of a test may sometimes be noted when, in the course of the research, test papers are made available to the psychiatrist for interview purposes. However much psychiatrists may assert that their opinions are uninfluenced by scores, there always remains an open question as to whether their criterion judgments have really remained uncontaminated despite knowledge of the variable whose validity is being assessed.

A third research procedure which too easily leads to unjustifiable optimism involves the use of an experimental population for the development of a scoring key and the immediate application of the key to the same population. Explanatory and apologetic footnotes may serve as a balm to the experimenter's conscience but they rarely serve as a guide to the kind of attenuation to be anticipated when the test and key are applied to a new population.

Ideally, tests to be validated would be administered to a population of inductees, the total group would then be admitted to military service and follow-up studies at various intervals would be made to determine success or failure of adaptation to the demands of military life. Every man tested would be given a very careful, extensive, and independent psychiatric examination by the most competent medical personnel available. In practice, however, theoretical perfection was necessarily sacrificed to operational schedules and personnel shortages. Typical psychiatric examinations were brief and generally devoid of careful preliminary elaboration of case histories. Fortunately, considerable work has been done in the Navy in condensing the psychiatric interview and yet preserving its essential values. It is felt that, in most instances, the criteria available possessed a high degree of significance as measures of test validity.

Findings

It is possible to present only major findings of a few sample studies conducted with some of the instruments previously described.[3] These

[3] Low scores in these instruments imply good adjustment and high scores imply poor adjustment.

will serve to illustrate the general research design employed as well as typical results obtained. Table 1-IX summarizes the procedures followed in connection with three different projects aimed at assessing and comparing the validity of four different screening tests.[4]

Research on Enlisted Personal Inventory, Form 2

This instrument was included in both Study I and Study II summarized in Table 1. Direct comparison of the findings in these two studies is difficult since different keys were employed in each research and the criteria are comparable only in a general sense.

STUDY I. In this study 561 men were tested and then given independent examinations and ratings by both a clinical psychologist and a psychiatrist. Since the data were accumulated over a period of time, the first 340 cases received were analyzed independently as Sample 1, and the remaining 221 were employed as a second or check sample—Sample 2.

Each subject was rated by both a psychiatrist and a psychologist on a scale defined as follows:

Group I. The group which can function with reasonable effectiveness in military service even though minor, non-disabling psychological symptoms may be present (treated as normal men in this study).

Group II. The group with marked psychological symptoms of such a character that there exists a doubt about their successful adjustment to the military service though no action, administrative or otherwise, is presently indicated (treated as maladjusted men in this study).

Group III-A. The relatively small group of men whose psychological adjustment is so poor that they are considered as temporarily disabled for effective military service and whose condition would warrant their being considered for such administrative disposition as: shore duty; extended leave; special assignment; rest camp; etc.

Group III-B. Those few cases who are so seriously disabled in the psychological sense as to warrant hospitalization or immediate survey (Both A and B were treated in the maladjusted group in this study).

In scoring the Enlisted Personal Inventory, Parts 1 and 2 were considered as separate tests. Part 1 was scored with a 20-item key

[4] The assistance of Lt. R. B. Porter and Lt. W. A. Owens in planning and conducting Study I, and of Lt. Comdr. G. A. Zirkle in planning and conducting Study II is acknowledged.

TABLE 1-IX. Description of research projects in the validation of psychiatric screening tests.

Study	Naval Installation	Population	Tests	Procedure	Criteria
I	Receiving Station A (Two samples)	561 enlisted men, most of whom were from the European Theatre, about half of whom had combat experience.	Enlisted Personal Inventory Personal Check List	Men were selected at random and given both tests. All men tested were interviewed separately by both a clinical psychologist and a psychiatrist. No test scores were available to the interviewers. Each man was given an independent rating by each interviewer on a scale from 1 to 3. Rating definitions are reproduced along with findings.	Ratings by psychiatrist and clinical psychologist.
II	Naval Training Center B (Two samples)	4,303 enlisted men entering recruit training. These were almost entirely raw recruits without previous military experience.	Enlisted Personal Inventory Billet Qualifications Blank	Men were selected at random and given both tests. Tests were then locked away unscored. A check was made at the Psychiatric Unit about three months after testing to determine which men had been referred for psychiatric investigation and what disposition had been made of the cases.	Administrative disposal based upon psychiatric judgment.
III	Receiving Station C	582 enlisted men, all of whom had overseas duty, being processed for reassignment.	Billet Qualifications Blank Experience Comparison Index	Men were interviewed by either a psychiatrist or a clinical psychologist and rated according to an agreed scale from 1 to 4. Rating definitions are reproduced along with findings. Testing took place after interviews and ratings were completed.	Ratings by psychiatrist and clinical psychologist.

developed by previous research.[5] Part 2 was scored on all 32 items without reference to the special significance of stop items. This crude procedure was followed because of some earlier indications given by the psychologists at Receiving Station A that such scoring had proved effective. Since an item analysis was not attempted, no information is available on the possibility of improving prediction by means of a key based on use of the more differentiating items.

As has been previously noted, two separate criteria were employed, namely, the ratings given by psychiatrists and the ratings given by clinical psychologists. This makes possible separate analyses for each type of criterion or for various combinations of criterion judgments. Such detailed data are available in the original research report but are too voluminous to present in this summary. For illustrative purposes, that portion of the data is presented which employs, as the criterion for maladjustment, the decision of *either* the psychiatrist *or* the psychologist that the subject belongs in rating groups II, III-A, or III-B. All other subjects are then automatically considered as belonging in group I, they being in effect the men considered by all examiners as normal.

Figure 1-IX presents a cumulative percentage distribution of scores on Part 1 for the normal and maladjusted groups in Sample 1, and in addition, the test score distribution for the total population in the sample. It should be noted that this type of graphic presentation permits a simple and rapid determination of the kind of prediction to be anticipated when any given percentage of the total group is referred for examination. If it is decided to refer 20% of the men for psychiatric investigation, then it is possible to read across from the 20% point on the vertical axis to the line representing the total population. On the base line will then be seen the test score which should be used to secure referral of approximately the desired percentage. At the same time it is possible to determine what the research has indicated in terms of the percentage of the normals who will be included in the referred group as well as the percentage of the maladjusted men who will be included. In view of the fact that these tests are not designed for use with specific cutting scores but are for use in referral of varying percentages of men depending on numbers that can be conveniently handled by psychiatrists at a given time, the practicality of this type of graphic presentation can readily be seen.

One special distribution curve contained in Figure 1-IX requires

[5] In general, item selection in these researches was primarily predicated on the effectiveness with which the item differentiated between criterion groups (e.g., maladjusted vs. normal). Inter-item correlation, item difficulty, and selection for internal consistency have been considered in connection with some item analyses.

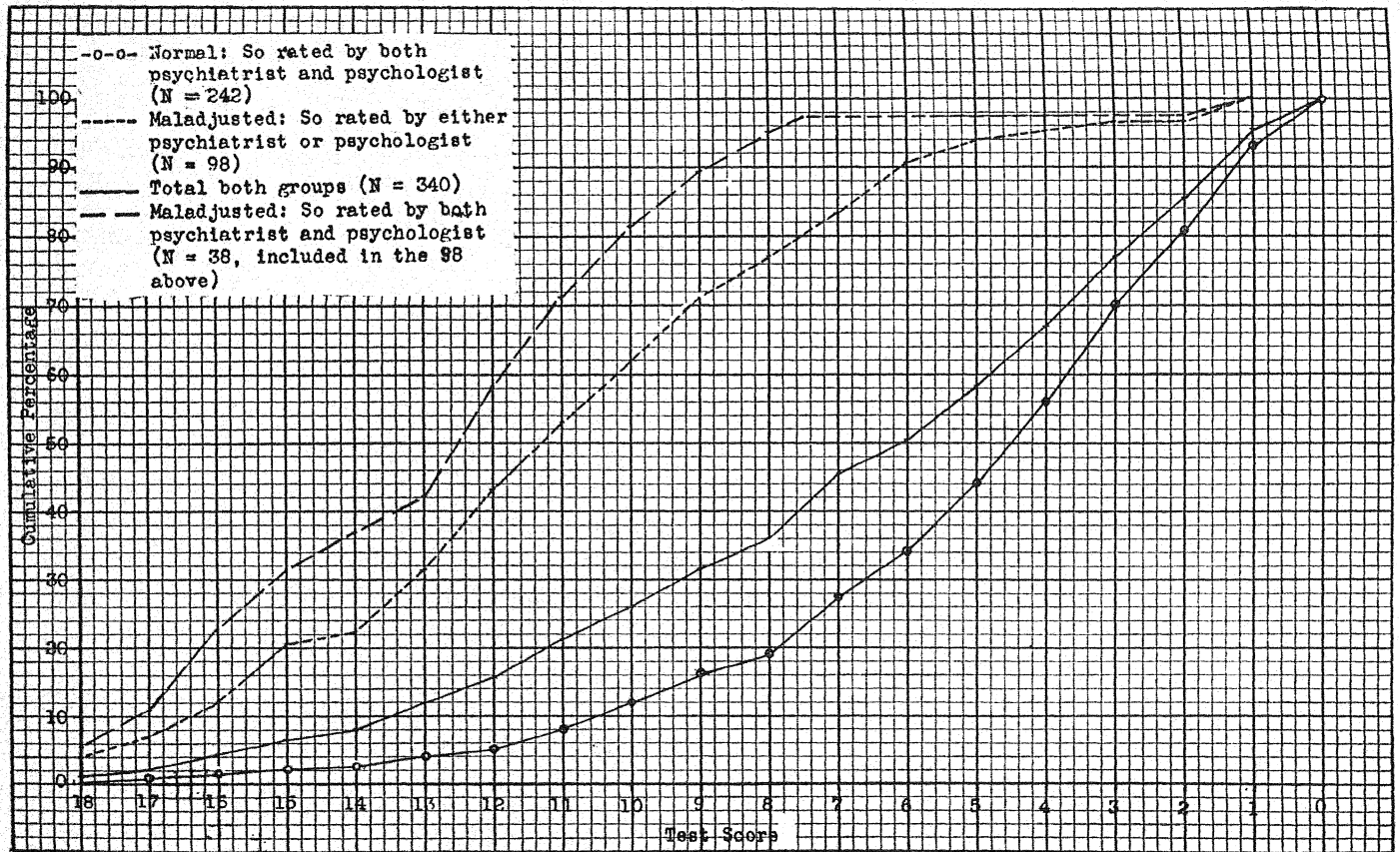


Figure 1-IX. Enlisted Personal Inventory, Form 2, Part 1: Percentage of normal and maladjusted men in Sample 1 tested at Receiving Station A who, at various score levels, would be referred for psychiatric examination. The figures in this chapter are read as follows: If it is desired to refer 50 per cent of the population for psychiatric interview, a cutting score of 6 should be set in the test. At this cutting score 90 per cent of the maladjusted (90 percent of 98) and 34 per cent (34 percent of 242) of the normals will be referred.

individual explanation. It will be noted that two separate distribution lines have been drawn for maladjusted groups. One of these (N=98) is complementary to the curve for the normal population and together with this latter group produces the curve for the total population. A second distribution curve (N=38) represents only a portion of the cases included in the larger maladjusted group and was selected out of the larger group in order to illustrate a point previously made in the discussion of the criterion. Whereas a subject might have been included in the larger maladjusted group because of the diagnosis of *either* the psychiatrist *or* the psychologist, only those men who were called maladjusted by *both* examiners were included in the smaller group. Presumably this latter group represents the most deviated population, clearly identifiable as such on two separate examinations. It is interesting to observe, therefore, that the cumulative percentage curve for this small group is markedly higher than for any other group. Expressed another way, it means that for this group, in which the most stringent criterion of independent identification by two separate examiners is employed, test prediction is easiest. Repeated illustrations of this point, that the more stringent the criterion the easier the prediction, could be presented in all the remaining graphs; but it will suffice to say that in this regard Figure 1-IX offers a typical result.

Figure 2-IX presents data on the same men (Sample 1) but this time for Part 2 of the Enlisted Personal Inventory. These data are directly comparable with those contained in Figure 1-IX, since the same criterion is employed in both instances. To simplify the comparison of these data, Table 2-IX indicates the discriminative power of Parts 1 and 2 of the Enlisted Personal Inventory at successive percentage levels of referral. Further comparisons of a similar nature for other referral points may be read directly from Figures 1-IX and 2-IX.

Simple inspection is sufficient to indicate that Parts 1 and 2 of the test have functioned with approximately equal effectiveness in predicting the deviated group and at approximately the same cost in the inclusion of false positives (i.e. normals incorrectly referred on a test score basis). Whatever differences do occur are too slight to invite positive interpretation.

Results for Sample 2 are similar to those in Sample 1 and do not require separate presentation.

Since Part 1 of the Personal Inventory contains items in forced-choice form whereas Part 2 items are cast in yes-no format, it may appear that the form of the item does not influence the validity of the items or that of the total test. Such an interpretation fails to take into account the possibility that Part 2 items may cover a more detailed and significant clinical symptom picture than Part 1 and that

this more adequate coverage compensates for the possibly less effective item format used in Part 2. Indeed, a reading of the two parts suggests that the item format of Part 1 is intrinsically sounder, but that the symptom inquiry in Part 2 is clinically more profound. As

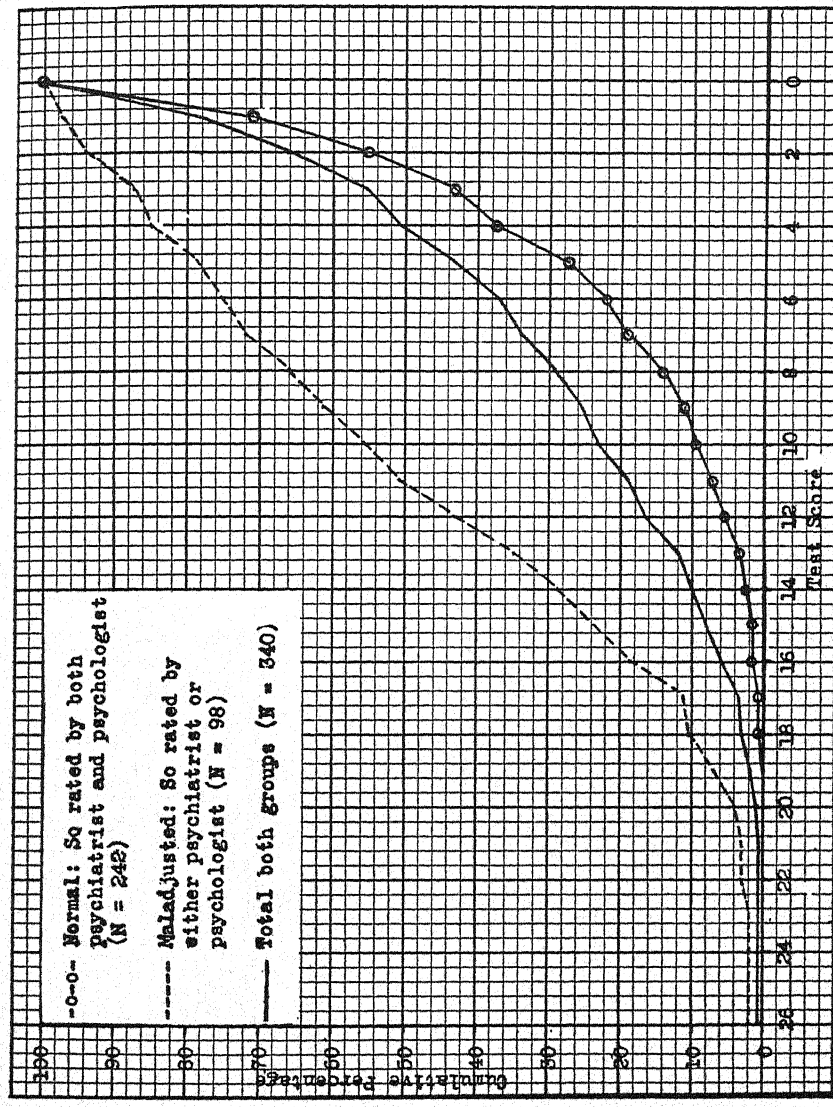


Figure 2-IX. Enlisted Personal Inventory, Form 2, Part 2:. Percentage of normal and maladjusted men in Sample 1 tested at Receiving Station A who, at various score levels, would be referred for psychiatric examination.

yet there is no conclusive answer to the very important question as to the most effective item structure. One independent investigation, indirectly related to the present research but conducted with other test instruments, produced data tending to support the notion that

the paired-choice type of item is substantially superior to the yes-no type. Until items of similar content but different format are tested against the same criterion it will be impossible to arrive at a definite decision in the matter. The question is of sufficient importance, however, to warrant further intensive study.

STUDY II. Use of the Enlisted Personal Inventory at Naval Training Center B differed from the procedure at Receiving Station A in that no special psychiatric investigations were initiated as part of the research study. The regular screening process was permitted to operate at Naval Training Center B and the development of a criterion came as a result of a follow-up inquiry at the Psychiatric Unit

TABLE 2-IX. Enlisted Personal Inventory, Parts 1 and 2: Percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Receiving Station A, Sample 1. (Percentages obtained from Figures 1 and 2)

Percentage of total population to be referred on test score basis	Percentage of normal men who would be incorrectly referred by Enlisted Personal Inventory		Percentage of maladjusted men who would be correctly referred by Enlisted Personal Inventory	
	Part 1	Part 2	Part 1	Part 2
10%	3%	3%	27%	29%
15	5	5	40	39
20	7	7	51	51
25	11	11	60	61
30	14	15	69	67
40	23	25	80	76

NOTE.—This table and similar ones which follow are interpreted thus: If 10 per cent of the total population is to be referred for psychiatric examination, Part 1 of the Enlisted Personal Inventory will incorrectly refer 3 per cent of the normals and correctly refer 27 per cent of the maladjusted. Comparable percentages for Part 2 are 3 and 29 respectively.

made some months after the data had been collected. A tabulation was then made with respect to the men tested to determine which ones had been referred for investigation as potential disability cases and which of the men had actually been discharged on neuropsychiatric grounds.

The following division was made of the total population of 4,303 men who had been tested:

Group Ia —1,996 normals (accepted for duty). Tests used in item analysis.

Group Ib —1,996 normals (accepted for duty). Used in prediction study.

Group IIa —58 referrals returned to duty. Tests used in item analysis.

Group IIb —57 referrals returned to duty. Used in prediction study.

Group IIIa—97 dischargees (neuropsychiatric). Tests used in item analysis.

Group IIIb—99 dischargees (neuropsychiatric). Used in prediction study.

This division of the population permitted item analysis employing normal and deviated populations and subsequent cross validation of the scoring keys on groups not involved in the item analysis.

TABLE 3-IX. Enlisted Personal Inventory, Form 2, Parts 1 and 2: Percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Naval Training Center B, Groups Ib, IIb, IIIb. (Percentages obtained from Figures 3, 4, and 5)

Percentage of total population to be referred on test score basis	Percentage of normal[1] men who would be incorrectly referred by Enlisted Personal Inventory			Percentage of maladjusted[2] men who would be correctly referred by Enlisted Personal Inventory		
	Part 1	Part 2	Parts 1 & 2	Part 1	Part 2	Parts 1 & 2
10%	7%	7%	7%	55%	57%	62%
15	12	12	12	67	64	70
20	16	17	16	72	71	75
25	21	21	20	78	74	79
30	26	26	26	79	78	81
40	36	35	37	82	85	85

[1] Group Ib: Well adjusted.
[2] Group IIIb: Discharged.

Figures 3-IX, 4-IX, and 5-IX present the cumulative percentage distributions for the Personal Inventory, Part 1, Part 2, and Parts 1 and 2 combined. Only those cases not used in the item analysis were included in these distributions.

Item analysis had resulted in the development of a 17-item key for Part 1, three of the original items having been found to have insufficient validity for retention. Nineteen items were retained in the key for Part 2. No effort had been made to eliminate overlapping items in the two parts since each was to be employed separately. The combination of scores was accomplished simply by adding the scores on the separate parts.

Table 3-IX summarizes results at various possible percentage levels

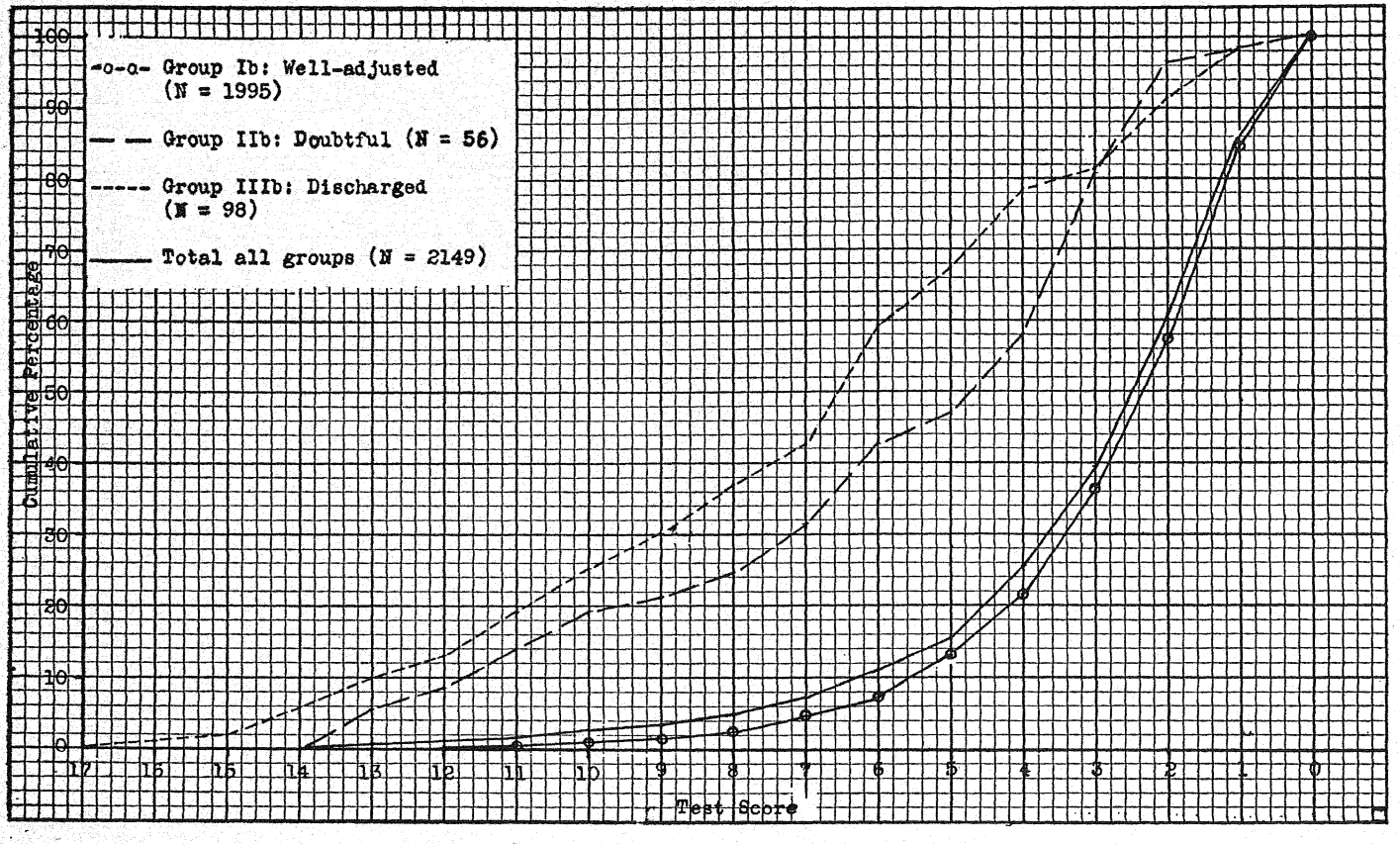


Figure 3-IX. Enlisted Personal Inventory, Form 2, Part 1: Percentage of men (well-adjusted, doubtful, and discharged) in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination.

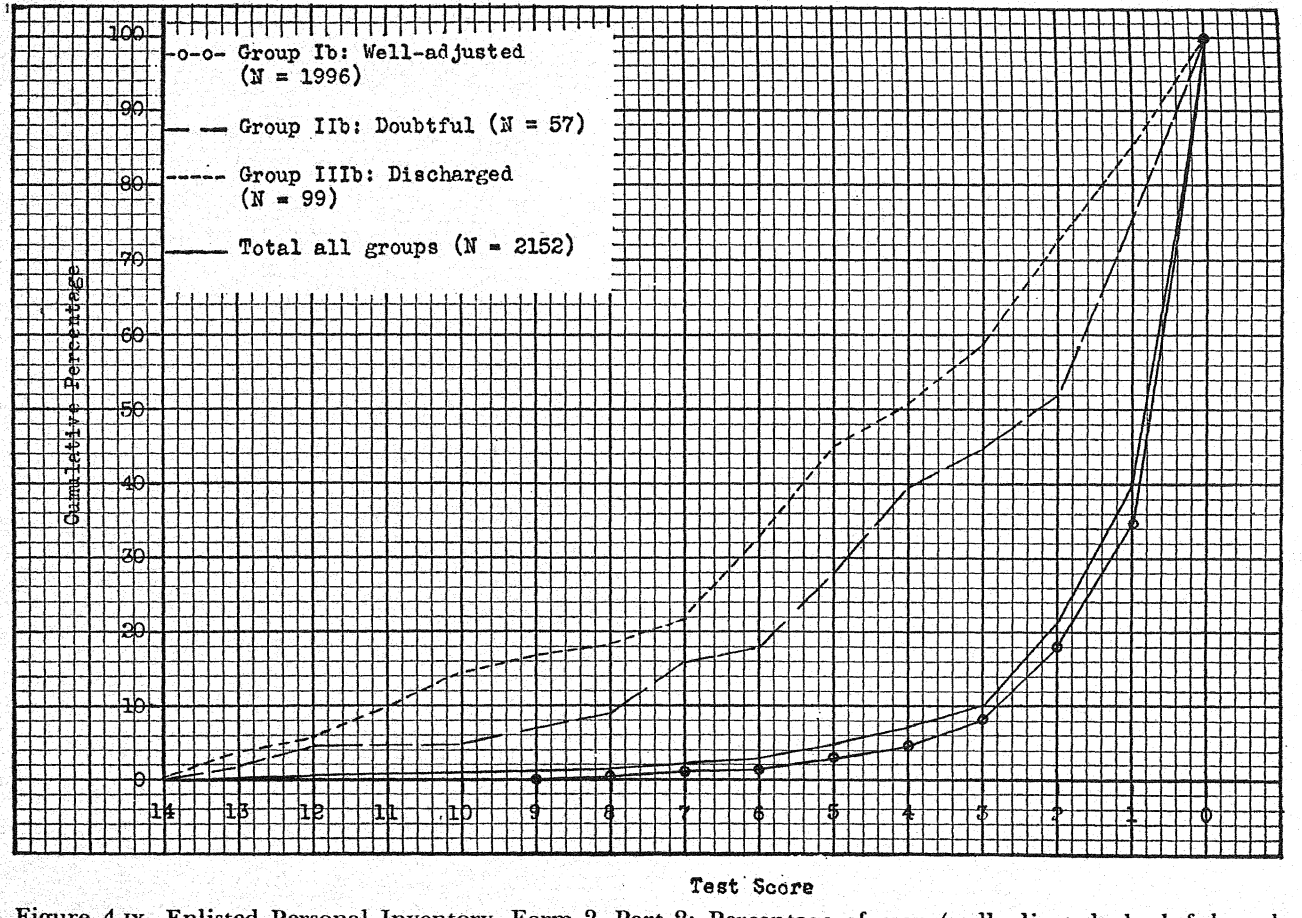


Figure 4-IX. Enlisted Personal Inventory, Form 2, Part 2: Percentage of men (well-adjusted, doubtful, and discharged) in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination.

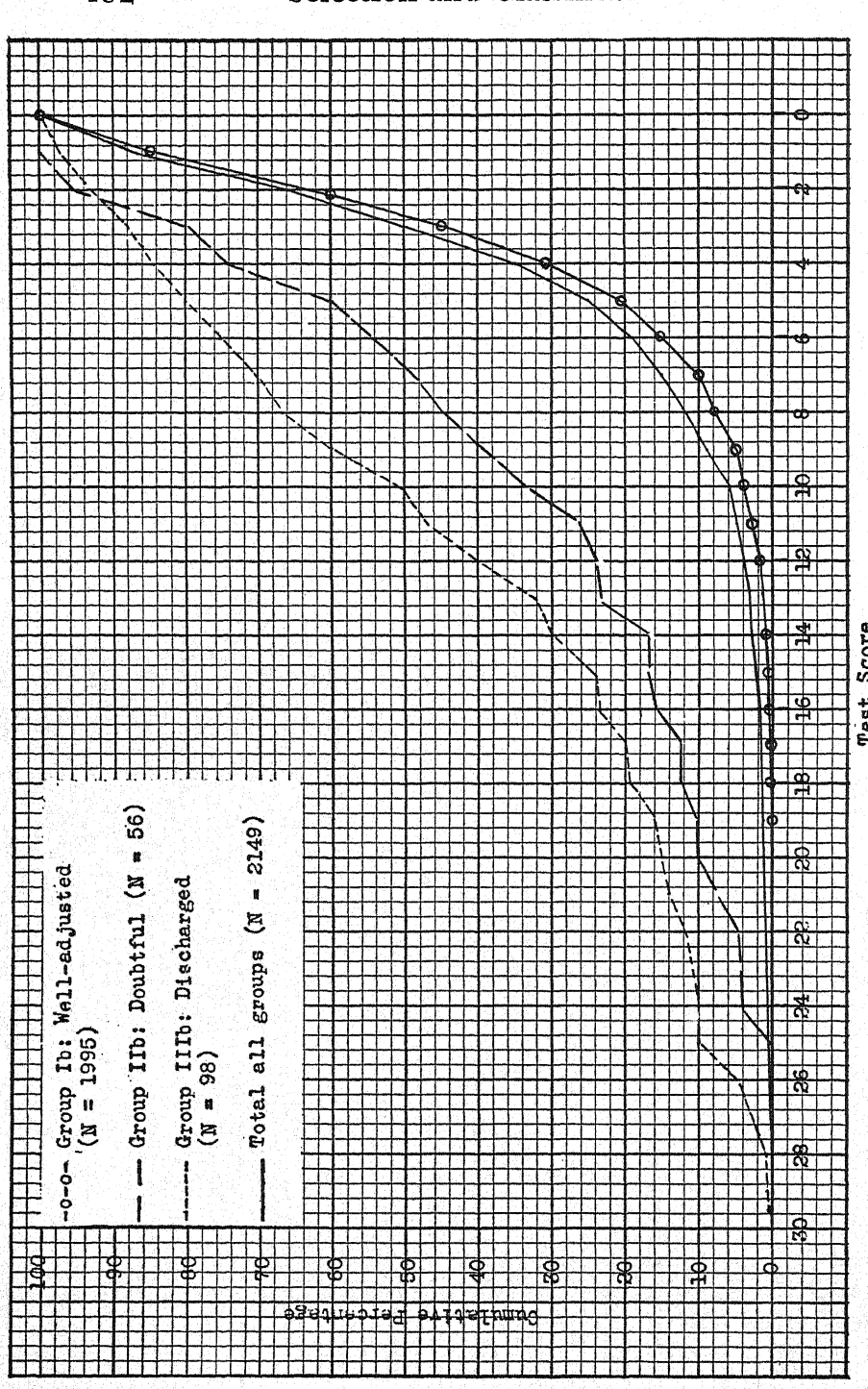


Figure 5-rx. Enlisted Personal Inventory, Form 2, Part 1 plus Part 2: Percentage of men (well-adjusted, doubtful, and discharged) in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination.

of referral, enabling immediate comparison of the graphs at a few specific points. As previously indicated, further comparisons are possible by direct reference to the graphs.

The results obtained in Study II are in agreement with those obtained in Study I, at least to the extent of establishing that Parts 1 and 2 of the Enlisted Personal Inventory have approximately equal validity. Furthermore, it would appear that the combination of part scores adds nothing to the predictive power of either part. This is in harmony with repeated findings that a relatively few items dealing with symptomatology operate just as effectively as a long and elaborate collection of items. It seems probable that, with the symptom measurement approach, the use of more than 20 or 30 items adds nothing to validity of prediction.

For the recruit population the intercorrelation of Parts 1 and 2 was .80. This supports the previous finding that these measures are interchangeable as they now stand. The relationship between scores on either part and scores on the General Classification Test (intelligence) is low, with correlations ranging from —.02 to —.13.[6] It should not be supposed from such correlations that intelligence is a negligible factor in the psychiatric screening process. Actually the mean General Classification Test score of the discharge group is more than half a standard deviation below the mean for the total population. Intelligence, then, must be considered a significant factor in the psychiatric screening process even though its correlation with scores on the personality inventories is fairly negligible.

Research on Billet Qualifications Blank, Form X-2(M)

STUDY II. At the same time that the men were given the Enlisted Personal Inventory at Naval Training Center B, they were also administered the Billet Qualifications Blank. The population is identical with that discussed in the immediately preceding section, and all research procedures presented there were duplicated exactly.

As a result of item analysis, that segment of the Billet Qualifications Blank which depended on the measurement of personal beliefs and social attitudes as a neurotic index was discarded. A single 39-item predicting scale, hereafter called the N-Scale, was developed for the measurement of maladjustment. The 10 items of the self-idealization scale, hereafter called L-Scale, were reduced to a scoring key with 4 items. These few items were chosen in terms of the

[6] Since high scores on the Personal Inventory are in the direction of maladjustment, these correlation coefficients generally appear with the negative sign after computation. The coefficients must actually be interpreted to mean that there is a slight tendency for individuals who score as maladjusted on the inventory to score lower on the test of intelligence.

theoretical premise previously discussed, namely that a suppressor variable should have a low or zero correlation coefficient with the criterion and be substantially correlated with the prediction variable. To the extent that the L-Scale conformed to this standard, it might be expected to operate in the direction of eliminating from N-Scale measurement those factors which were irrelevant to the prediction of the criterion.

Figure 6-IX presents the cumulative percentage distribution of N-Scale scores for the men in Groups Ib, IIb, and IIIb at Naval Training Center B. Figure 7-IX presents cumulative percentage distributions of scores for the same men when L- and N-Scale scores were combined. The significant question with respect to these data

TABLE 4-IX. Billet Qualifications Blank, Form X-2(M), Scale N and Scale N plus Scale L: Percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Naval Training Center B, Groups Ib, IIb, IIIb. (Percentages obtained from Figures 6 and 7)

Percentage of total population to be referred on test score basis	Percentage of normal[1] men who would be incorrectly referred by Billet Qualifications Blank		Percentage of maladjusted[2] men who would be correctly referred by Billet Qualifications Blank	
	Scale N	Scales N + L	Scale N	Scales N + L
10%	7%	7%	49%	54%
15	12	11	64	64
20	17	17	67	69
25	21	21	71	72
30	26	26	75	77
40	36	36	81	84

[1] Group Ib: Well adjusted.
[2] Group IIIb: Discharged.

is whether the 4-item self-idealization key has made any improvement in prediction, whether in fact this modest suppression key has actually corrected the scores on the maladjustment scale in the proper direction.

Table 4-IX shows the results for Naval Training Center B for the N-Scale alone and for Scales N plus L. These data are so presented that they may be compared with the referrals made by the Enlisted Personal Inventory as shown in Table 3-IX.

There is a slight consistent improvement in prediction when the suppression key is used along with the key for the measurement of maladjustment. The improvement is small and can only be taken as indicating that the technique of assessing an idealization trend

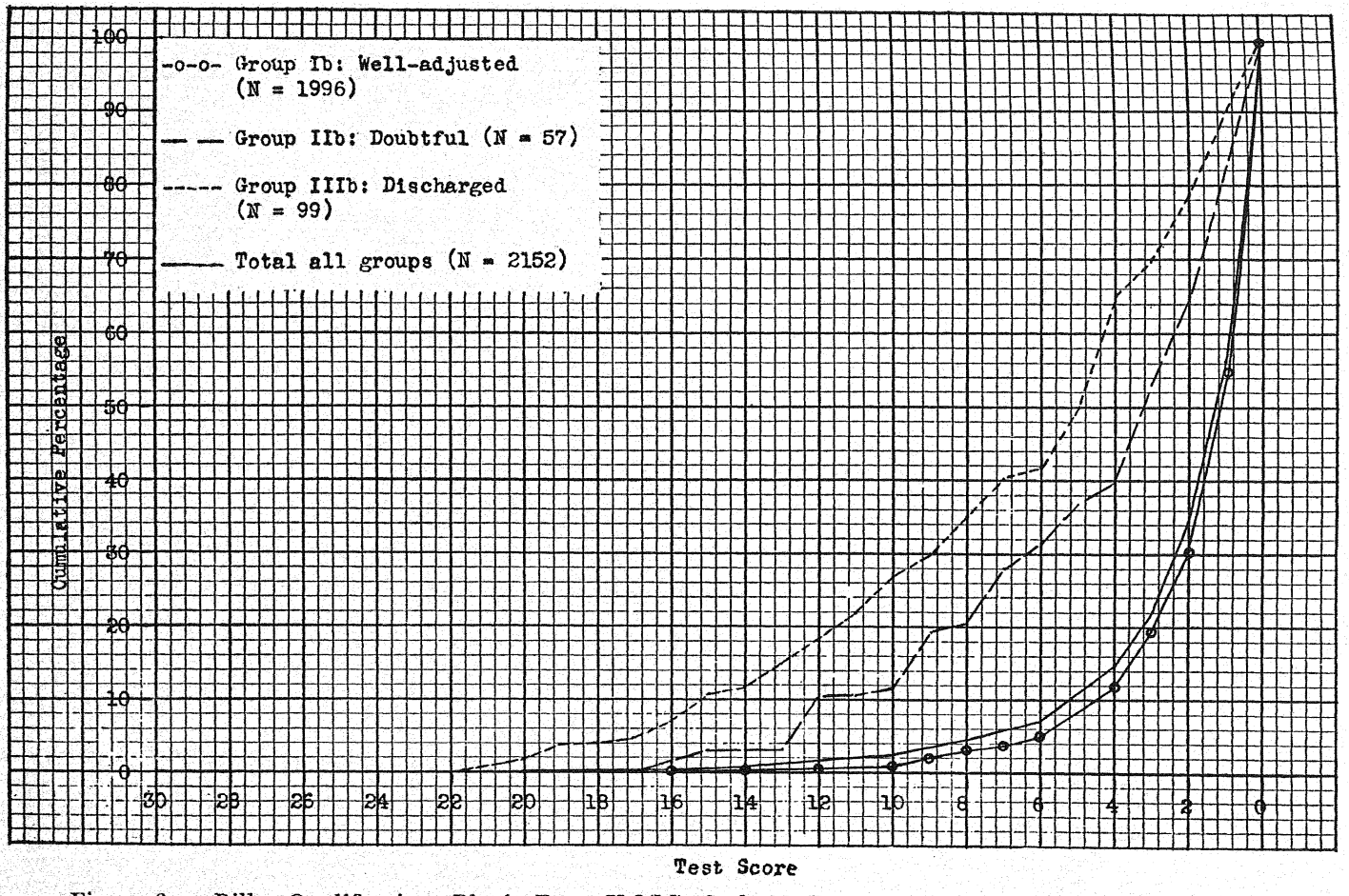


Figure 6-IX. Billet Qualifications Blank, Form X-2(M), Scale N: Percentage of men (well-adjusted, doubtful, and discharged) in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination.

in order to correct a score for maladjustment is worth further investigation.

The Billet Qualifications Blank as a whole differentiated the discharged population from the accepted population somewhat less sharply than the Personal Inventory, though the differences are too slight at any referral level to have much statistical significance. Whether this is due to the fact that a billet assignment context for the questions has no important effect on test validity, or because

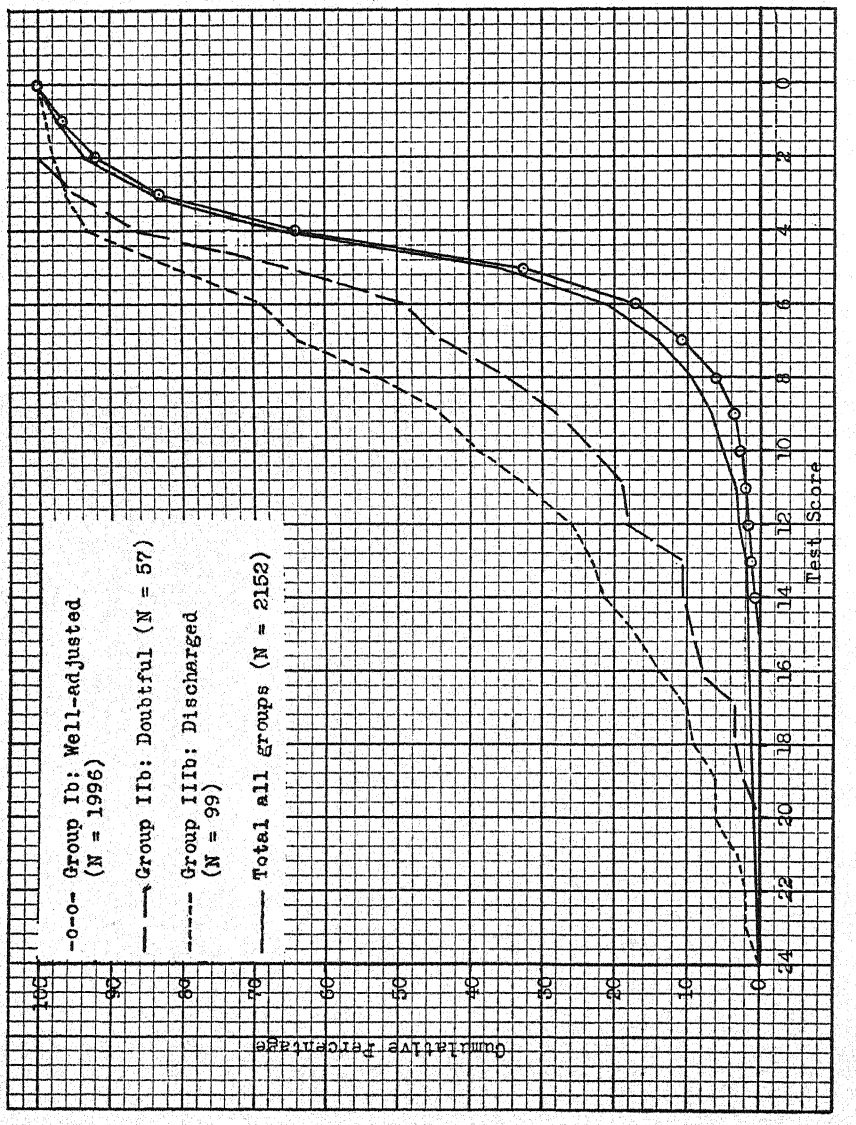


Figure 7-IX. Billet Qualifications Blank, Form X-2(M), Scales N plus L: Percentage of men (well-adjusted, doubtful, and discharged) in Groups Ib, IIb, and IIIb tested at Naval Training Center B who, at various score levels, would be referred for psychiatric examination.

the item contents of the two tests are of unequal clinical significance, or because the item formats differed in other respects than the billet context cannot be determined at this time. It seems improbable however that any large improvement in prediction can be anticipated from linking the inventory with the billet problem.

STUDY III. Validation of the Billet Qualifications Blank at Receiving Station C proceeded on a somewhat different basis than at Naval Training Center B. Whereas in the latter station the routine screening process provided the means for establishment of a criterion, special psychiatric and psychological examinations were instituted at Receiving Station C in connection with the entire experimental population. These examinations were conducted before testing took place and the psychologists and psychiatrists who cooperated in the project employed a pre-established rating scale for evaluating each subject.

The specific definitions employed for rating the men were as follows:

Group I Fit for sea duty; shows no disabling features of a psychological nature; despite sea or foreign shore duty, shows no obvious combat or operational fatigue symptoms; looks like A-1 material; should not break down under stress (used in study as normal group).

Group II Probably fit for sea duty; shows minor disabling features of a psychological nature; some combat fatigue, operational fatigue, or nervous tension symptoms, but may be able to weather further sea or foreign shore duty (used in study as doubtful group).

Group III Questionable material; shows disabling features of a psychological nature; tension or fatigue symptoms evident; may or may not be useful aboard ship (poor adjustment group in study).

Group IV Unfit for sea duty; marked disability and clearly unsuitable for further active service afloat at this time (poor adjustment group in study).

Groups III and IV were combined in the study because: (1) there were only nine men allocated to Group IV and (2) the examiners who had done the screening gave indications that men allocated to Group III were distinctly inferior in adjustment to the problematic cases contained in Group II. It should be noted that whereas over 600 men were tested and rated in this study, a lesser number will be reported both in connection with the Billet Qualifications Blank and with the Experience Comparison Index (which was administered at the same time). This discrepancy is due to the fact that a small

segment of the population was eliminated from the prediction group after being used for item analysis purposes.

Figures 8-IX and 9-IX present the test score data for the Billet Qualifications Blank, N-Scale and Scales N plus L, on a sample of 268 cases.[7] These data from Receiving Station C are difficult to compare with those from Naval Training Center B. The criteria were different in terms of stringency, procedure, and personnel responsible for their development. The populations were very different, Naval Training Center B involving only recruits entering basic training, Receiving Station C involving men who had had extensive tours of sea duty. Even the scoring keys were different, the N-Scale consisting of 40 items which overlapped the 39 items em-

TABLE 5-IX. Billet Qualifications Blank, Form X-2(M), Scale N and Scale N plus Scale L: Percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Receiving Station C. (Percentages obtained from Figures 8 and 9)

Percentage of total population to be referred on test score basis	Percentage of normal[1] men who would be incorrectly referred by Billet Qualifications Blank		Percentage of maladjusted[2] men who would be correctly referred by Billet Qualifications Blank	
	Scale N	Scales N + L	Scale N	Scales N + L
10%	.5%	2%	25%	28%
15	3	3	35	38
20	4	4	46	52
25	6	6	53	60
30	9	8	63	67
40	13	13	75	78

[1] Group I: Well adjusted.
[2] Groups III and IV: Poorly adjusted.

ployed in Study II but differed in several substantial respects. Also, the L-Scale used in Study III employed all 10 items as contrasted with a restriction to 4 items used in Study II.

Table 5-IX offers a few comparative statistics on prediction by the N-Scale and by the N-Scale corrected through additions of L-Scale scores.

The interesting fact to be noted in the tabular and graphic data is that the suppression key has improved prediction noticeably. Though interpretation of this result must still be made cautiously, at least on the surface it would appear that the type of exaggeration

[7] For a second sample of 189 cases, treated in the same manner, similar results were obtained and hence are not presented here.

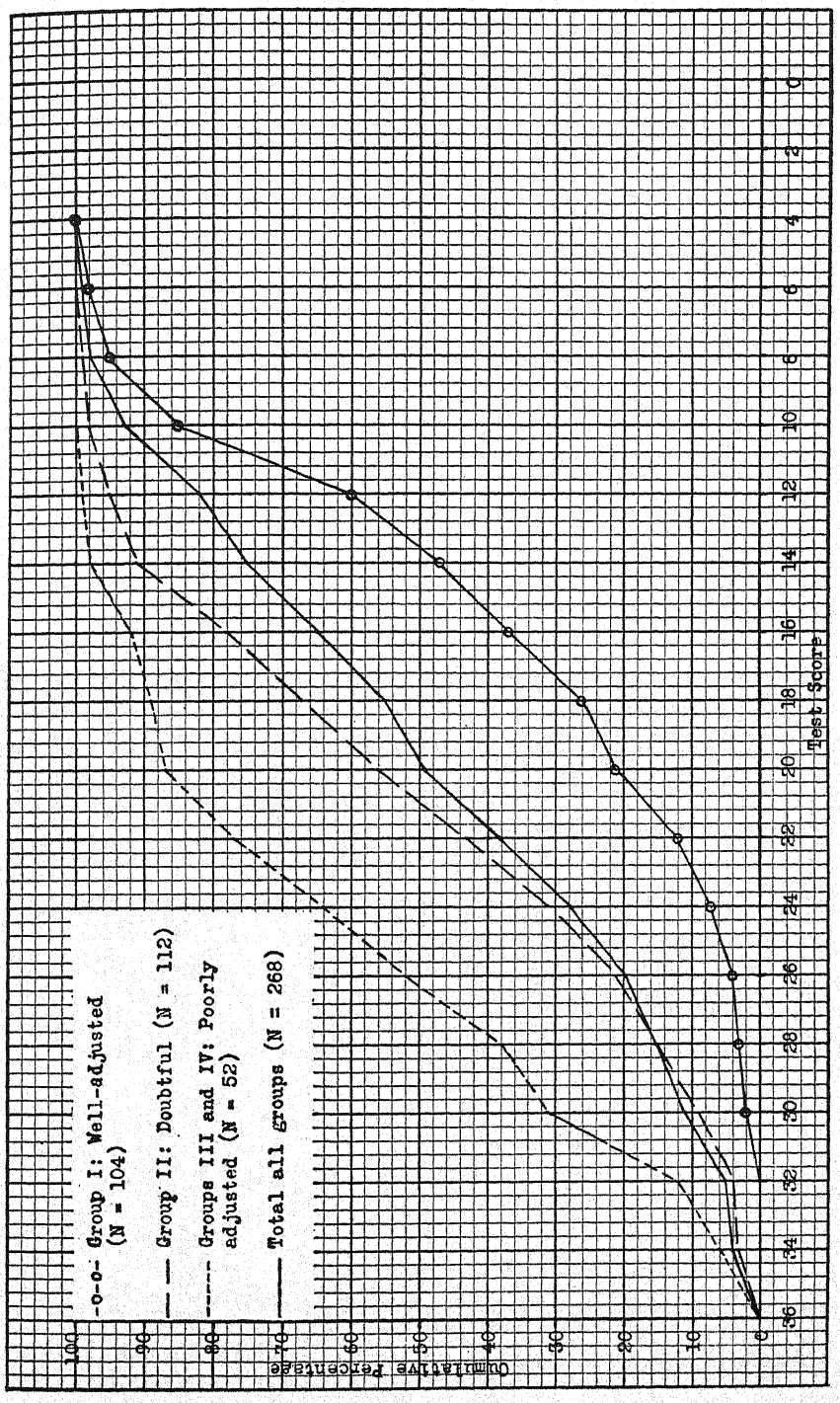


Figure 8-IX. Billet Qualifications Blank, Form X-2(M), Scale N: Percentage of men (well-adjusted, doubtful, and poorly adjusted) tested at Receiving Station C who, at various score levels, would be referred for psychiatric examination.

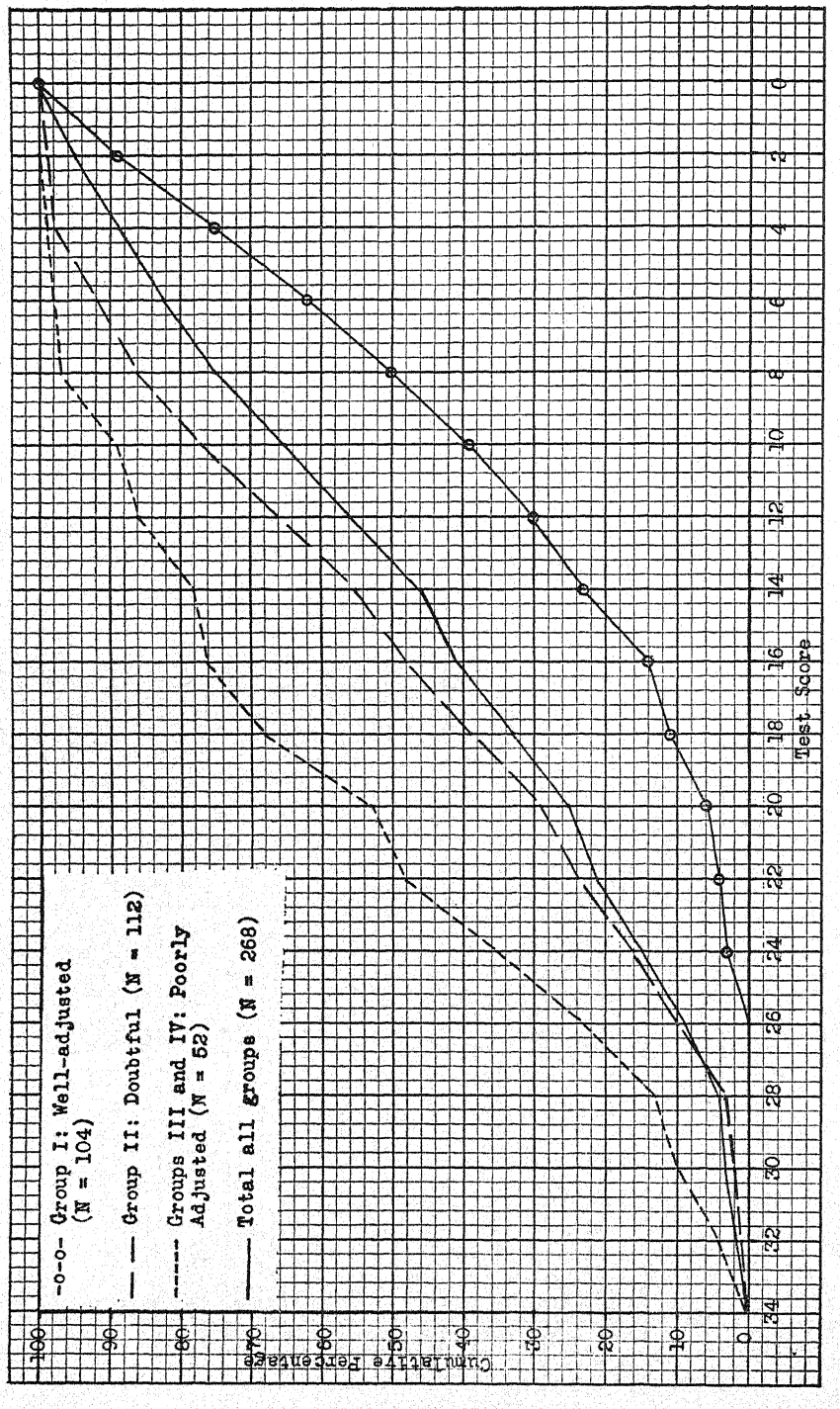


Figure 9-IX. Billet Qualifications Blank, Form X-2(M), Scales N plus L: Percentage of men (well-adjusted, doubtful, and poorly adjusted) tested at Receiving Station C who, at various score levels, would be referred for psychiatric examination.

measured by the L-Scale is actually a factor in distorting answers on the N-Scale and thus detracting from the prediction of the criterion. Adding the two scores together seems to have the effect of subtracting from the N-Scale scores an element which is uncorrelated with the criterion and thereby purifying such scores. Before such a conclusion would be warranted in fact, however, much better verification than is presently available would be needed.

Research on Experience Comparison Index, Form X-1

STUDY III. This test was administered to the same population as that used in the study of the Billet Qualifications Blank above. This enables direct comparison of the two tests for a situation which employed the same population as well as the same criterion.[8] Figure 10-IX, which presents the data for the Experience Comparison Index, includes a special distribution curve which must be considered separately from the remaining distributions. This curve represents the cumulative percentage of cases found at successive score levels for a population of hospitalized neuropsychiatric patients. As in the earlier presentation of a similar problem in Figure 1-IX, it will be noted that the more aberrant or deviated are the cases to be predicted (criterion of greater stringency), the more successful is the prediction. The hospitalized population obtains higher (therefore poorer) scores on these inventories than even a maladjusted group within a combat veteran population. If the contrast had been between the scores of hospitalized patients and a raw recruit population, the differences would have been even greater.

This finding has particular importance because of the frequency with which the literature on research in personality test validation recommends or uncritically supports the use of experimental groups with known psychological differences. Such differences must ultimately be established in any research if the validity of prediction is to be estimated. The danger lies in overestimating results when using groups already differentiated by previous psychiatric diagnosis, especially under conditions which have isolated only the most extreme deviates.

Table 6-IX permits ready comparison of the Billet Qualifications Blank and the Experience Comparison Index in Study III.

Prediction through use of the Billet Qualifications Blank is some-

[8] Differences in the sizes of populations reported for the Billet Qualifications Blank and the Experience Comparison Index are due to the fact that the analysis in connection with the former instrument split the total group (remaining after eliminating the item analysis group) in order to develop data on two separate samples. The analysis for the Experience Comparison Index employed a different number of cases for item analysis and then considered all the remaining cases as a single population.

what sharper than with the Experience Comparison Index. The similarities in the success of various instruments when used in situations which permit direct comparison are, however, far more significant than the differences which are noted. The general impression is gained that most of these inventories, regardless of item format or even the specific neurotic symptomatology investigated, produce about the same sharpness of prediction. It is almost as if a kind of plateau in prediction had been reached through the symptom ques-

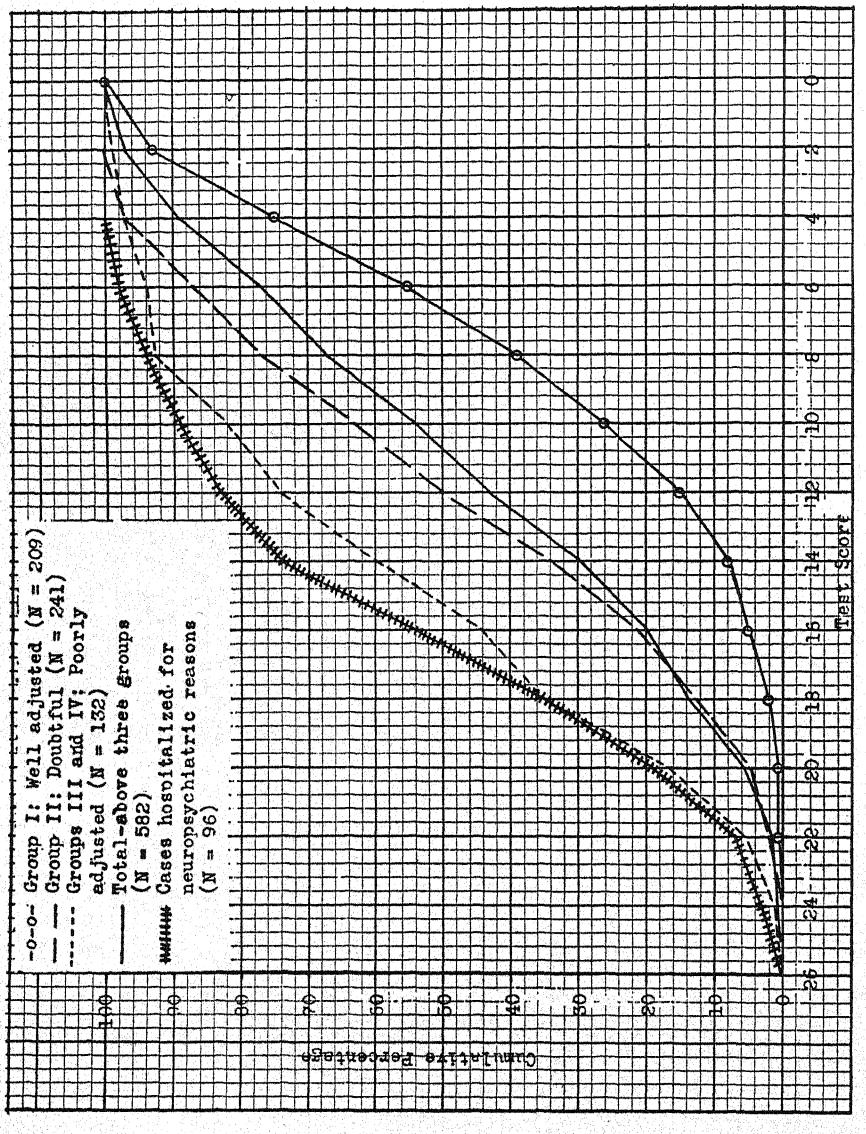


Figure 10-rx. Experience Comparison Index, Form X-1: Percentage of men (well-adjusted, doubtful, and poorly adjusted) tested at Receiving Station C and of neuropsychiatric cases, tested at a Naval Hospital who, at various score levels, would be referred for psychiatric examination.

tionnaire approach. Sharply different methods will probably have to be experimented with before very significant advances are made beyond the present accomplishments.

Research on Personal Check List, Form X-4

STUDY I. Data were presented earlier in this report on the use of the Enlisted Personal Inventory in predicting maladjustment among overseas veterans being processed at Receiving Station A.

Since the Enlisted Personal Inventory had been developed for use with recruits rather than with overseas veterans there was some question as to its validity with this latter group.

TABLE 6-IX. Experience Comparison Index, Form X-1, and Billet Qualifications Blank, X-2(M), Scale N plus Scale L: Percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population studied. Receiving Station C. (Percentages obtained from Figures 9 and 10)

Percentage of total population to be referred on test score basis	Percentage of normal[1] men who would be incorrectly referred by		Percentage of maladjusted[2] men who would be correctly referred by	
	Experience Comparison Index	Billet Qualifications Blank Scales N + L	Experience Comparison Index	Billet Qualifications Blank Scales N + L
10%	1%	2%	26%	28%
15	2	3	37	38
20	5	4	44	52
25	6	6	52	60
30	8	8	60	67
40	13	13	71	78

[1] Group I: Well adjusted.
[2] Groups III and IV: Poorly adjusted.

Preliminary experimentation at a West Coast receiving ship had resulted in the development of the Personal Check List, Form X-3. This 27-item instrument employed a number of items directly associated with the so-called "combat fatigue" syndrome as it is generally delineated. The general exploratory work involved in test construction had given some indication of the superiority of the new instrument over the Enlisted Personal Inventory when used with combat veterans. Study I was then set up as a more definite check on the comparative validity of the two tests for use with personnel possessing considerable operational or combat experience.

Before administering the Personal Check List at Receiving Station A, several experimental additions to the blank were made, though

no changes were made in the original 27 items. First, 15 relatively untried items dealing with symptomatology not covered by the earlier form were added to determine if they would improve prediction. Second, 15 items in the nature of a self-idealization scale were added. These latter items differed from those in the Billet Qualifications Blank in content and also in format, the paired-choice structure having been adopted in keeping with the rest of the Personal Check List. With these two additions, the X-3 form became known as Form X-4.

Figure 11-IX presents data on the first sample of Receiving Station A men tested by the Personal Check List, Form X-4. It will be noted that this represents analysis of the data for the same population as is treated in Figures 1-IX and 2-IX in which the Enlisted Personal Inventory distributions were presented. It should also be noted that Figure 11-IX covers results only for the 27-item key of the Personal Check List. This key included only the items originally contained in the earlier Form X-3 since it was soon apparent that the 15 additional symptomatic items and also the 15 items dealing with self-idealization trends had failed to produce any increase in discrimination for the test.

In view of the important theoretical problem involved, it would be well to examine more closely the fact that the 15 self-idealization questions had no marked effect either in the direction of increasing or decreasing the predictive efficiency of the test. Data are not available as to the correlation coefficients of individual items of the self-idealization scale with other items, with total test scores, or with the criterion. The analytic procedure used involved only the crude addition of scores on all 15 items in the self-idealization scale to scores on the total test. It is felt that the full potentialities of a suppression key have not thereby been adequately evaluated. It may very well be that the type of correlations found for the Billet Qualifications Blank will not be found in connection with an item format employing the forced-choice technique. But until more detailed analysis is completed, no final conclusions should be drawn.

Figure 12-IX also deals with the 27-item key of the Personal Check List but concerns itself with a second sample of 221 men at Receiving Station A.[9] It will be noted that with the Personal Check List prediction for this later sample is considerably sharper than for the first group. While it cannot be said with certainty which result more accurately assesses the predictive power of the instrument, there is some basis for relying on the findings obtained on analysis of the second sample population. This basis is found in the fact that dis-

[9] The second sample data were not previously presented for the Personal Inventory because results were so nearly like the findings for the first group.

tribution of all scores in the later experimental group closely approximates the distribution found for a very large population tested at a West Coast receiving ship. This was not true for the first and somewhat larger sample.

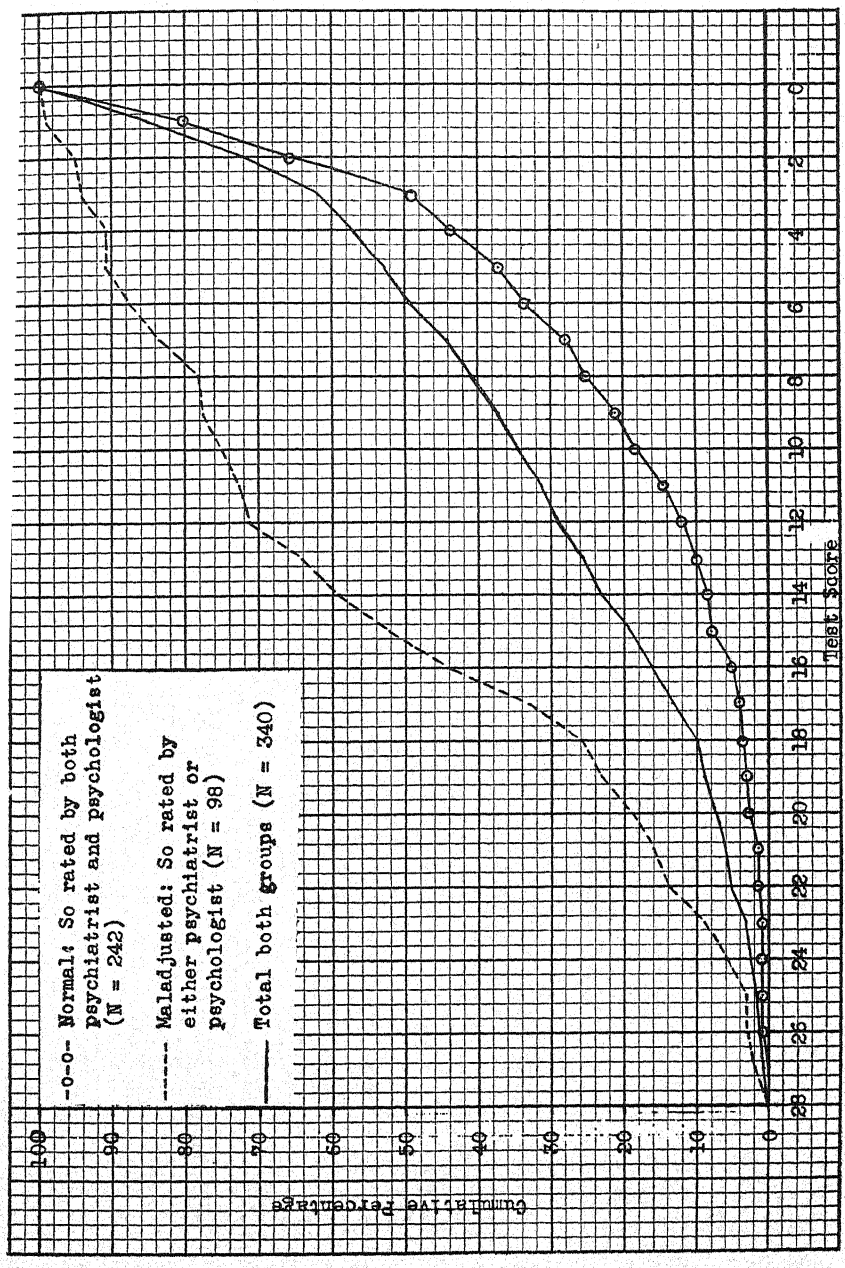


Figure 11-IX. Personal Check List, Form X-4 (27 Item Key): Percentage of normal and maladjusted men in Sample 1 tested at Receiving Station A who, at various score levels, would be referred for psychiatric examination.

In order to assess differences in validity as between the Personal Check List and the Enlisted Personal Inventory, and also to permit a comparison of results obtained with the Personal Check List on two separate samples of the Receiving Station A population, predictions at various referral levels are shown in Table 7-IX.

The comparative data on Sample 1 indicate that there is little to

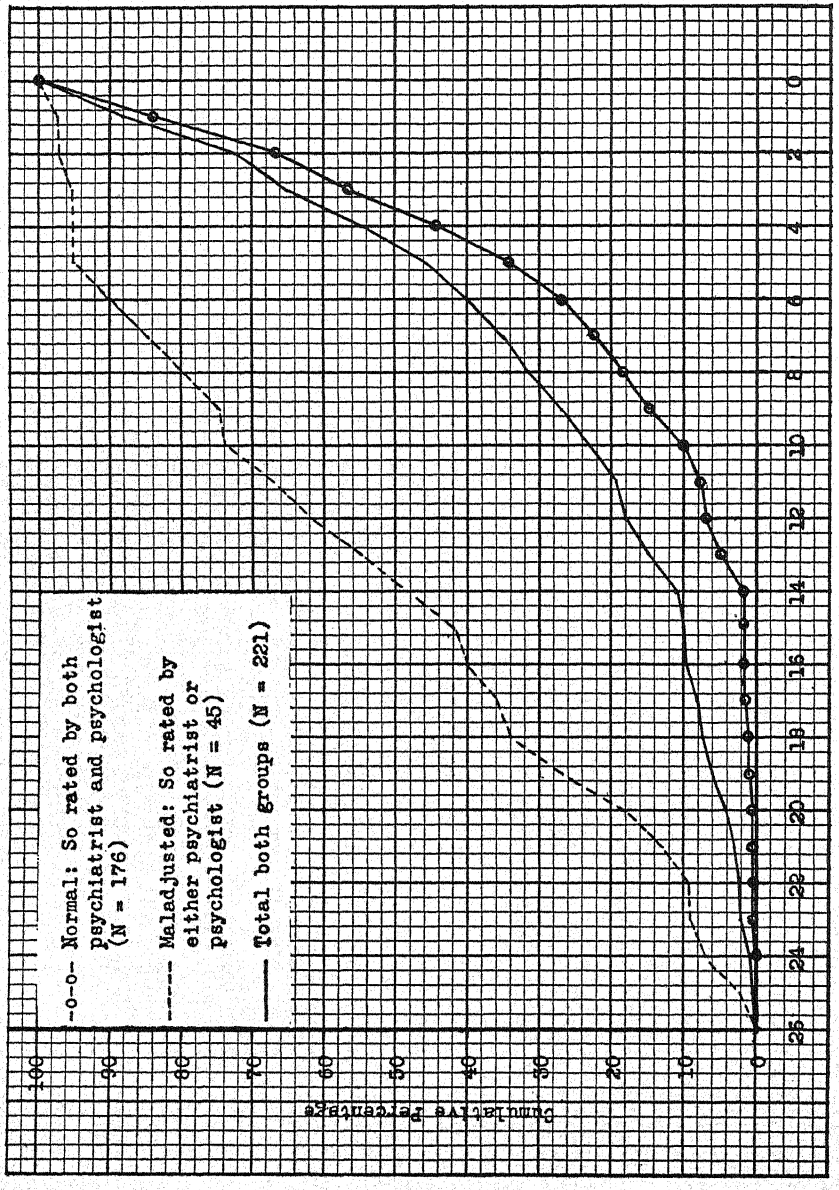


Figure 12-IX. Personal Check List, Form X-4 (27 Item Key): Percentage of men in Sample 2 tested at Receiving Station A who, at various score levels, would be referred for psychiatric examination.

choose between the Personal Check List and Part 1 of the Enlisted Personal Inventory. However, when the second sample is considered there can be observed a marked difference in predictive efficiency. The Personal Check List consistently picks up a higher percentage of maladjusted men, and though there seems to be some slight additional cost in false positives, the overall results definitely favor the Personal Check List. It is impossible to say whether a repetition of the experiment with very large samples would continue to produce the same results, but the evidence to date favors the Personal Check List at least for those personnel with considerable sea duty.

TABLE 7-IX. Personal Check List, Form X-4, and Enlisted Personal Inventory Form 2, Part 1: Percentage of normal and maladjusted men who would be referred for psychiatric examination at various percentage levels of total population tested. Receiving Station A, Samples 1 and 2. (Percentages obtained from Figures 2, 11, and 12)

Percentage of total population to be referred on test score basis	Percentage of normal men who would be incorrectly referred by			Percentage of maladjusted men who would be correctly referred by		
	Enlisted Personal Inventory	Personal Check List		Enlisted Personal Inventory	Personal Check List	
	Sample 1	Sample 1	Sample 2	Sample 1	Sample 1	Sample 2
10%	3%	3%	2%	27%	26%	44%
15	5	5	5	40	39	55
20	7	8	8	51	53	68
25	11	10	12	60	63	74
30	14	13	18	69	72	79
40	23	25	27	76	77	90

Other Studies

It has been necessary to omit a very considerable body of data dealing with other populations and instruments. A number of studies have been made using the inventories already reported and other tests in connection with applicants for submarine duty, with amphibious personnel, with midshipmen, with WAVES, with prison populations, etc. The implications of several findings drawn from these various studies deserve brief mention.

It has been interesting to note that a rather well defined hierarchy of means for any test can be found as different populations are tested. The lowest mean scores are generally found among submarine volunteers. Raw recruits have a somewhat higher mean. Amphibious personnel will generally have average scores that are even higher, and

combat veterans almost invariably obtain average scores higher than any of the other groups.

Such a state of affairs may be subject to several interpretations. It may be taken as evidence of validity in the sense that a test score hierarchy seems to fit general expectation concerning the adjustment of the populations in question. Submarine volunteers are highly selected and presumably well adjusted individuals. The raw recruit population is a relatively undifferentiated group. The amphibious personnel included a large number of "general detail" men who were not qualified or not needed for school or special assignments. Men who had combat experience might be expected to show the effects of such duty and exhibit symptoms of maladjustment to a fairly marked degree.

On the other hand, the situation is just as amenable to interpretation in terms of motivation. Men applying for submarine duty are highly motivated. As a volunteer group, they are extremely eager to make a good impression and qualify for submarine training. Raw recruits, too, generally have reasonably high motivation since all the indications are that men passing through the initial training stages are generally eager to do well in the service, though the motivation is hardly as high as in the case of the submarine group. There is reason to believe that the amphibious forces were somewhat less motivated (see Chapter XXI). Certainly the motivation of the combat-experienced groups, at least so far as the tests were concerned, was less than for any of the other groups. Simple observation has confirmed the fact that these men were indifferent to the tests and cynical concerning anything except the chance for immediate shore leave.

Under the circumstances it is impossible to elevate the test score hierarchy into proof of validity. It is probable that differences in actual adjustment as well as differences in motivation play some role in producing this picture. At all events it is clear that very different norms will be required for each of these varying populations. The effect of using a particular critical score for referral purposes in one group will not necessarily hold for another group. The score level which will accomplish the referral of only 20% of all recruits in a training center may be so low with respect to a receiving station processing combat veterans that 50% are referred, or so high with respect to submarine candidates that only 3% are referred.

The interesting picture then develops that a man who makes a certain score on a test at a recruit training center may, in terms of previous research, have an even chance of being discharged as a neuropsychiatric liability. On the other hand, a man who makes that same score at a receiving ship after combat experience may have

only 1 chance in 20 of being screened out. Perhaps this means that psychiatrists themselves adjust their thinking to the class of men with whom they deal. At training centers a certain symptom picture may be given considerable weight in bringing about the discharge of a man from naval service. The same symptom picture appearing in a man just returned from combat duty may be so commonplace and unimportant as to merit no more than passing attention. Or it may mean that the psychiatrists screen on some other basis than the mere summation of a series of symptoms as is true of the test. It is interesting to note, however, that even if a very different predictive value must be placed on a particular score in terms of the group from which it is derived, the test as a whole continues to have considerable validity for any of the populations on which it is used. All that seems to happen as one moves from a raw recruit population to a seasoned veteran population is that scores for both normals and maladjusted shift upward to about the same extent. Differentiation thereby remains at a fairly constant level.

Summary and Discussion

Whenever different inventories have been employed on identical populations and then evaluated in terms of the same criterion, it has been striking to note how similar have been the prediction results. In no instance has it appeared with clarity and certainty that a particular instrument, a particular content, or a particular format is decidedly superior to any other instrument, content, or format. It begins to appear as if psychiatric screening instruments employing symptom inquiry as the basic technique had reached a plateau in their development. It may be anticipated that refinement of this technique will produce small, though perhaps significant, improvements in prediction. Major advances, however, will probably depend on radical innovations in the methods of measurement.

Item Content in Relation to the Psychiatric Criterion. As a reasonably effective and economical instrument for psychiatric referral purposes, the symptom-oriented questionnaire has proved its value. It is therefore important to observe the particular items which seem consistently to offer the highest validities for prediction. A good many of the most significant items may be classified somewhat generally under the heading of conversion symptoms, or at least as having reference to the somatic representatives of anxiety and guilt. Almost without fail, the most valid single item in these inventories turns out to be one dealing with headaches. Its nearest rival, interestingly enough, is that type of item which quite straightforwardly inquires as to whether the subject considers himself nervous to any

considerable degree or has been treated or has contemplated treatment by a doctor for nervousness. Dizzy spells, fainting, tachycardia, stomach pains, physical debility, and similar psychosomatic complaints seem to hold up consistently on item analysis. Items which are less complicated by somatic overtones are also found to be discriminative but not to the same extent nor so unvaryingly as in the case of the conversion symptoms. In this class will be found questions dealing with dislike for being watched at work, phobic attitudes, difficulty in making friends or finding things to do in spare time, et cetera.

The situation is undoubtedly a reflection of the nature of the psychiatric screening process. After all, item analysis accomplishes a keying of the test to the psychiatric examination. Those items which most closely simulate the procedures and emphases of the psychiatrist's interview will generally turn up as the items of highest validity. In a sense it would be perfectly justifiable to say that the best way to build an inventory of the type sought after here would be to study the techniques of the psychiatrists rather than the disabilities of the subject. So long as the major object is to predict the psychiatric criterion, one should emphasize a study of the criterion. When the major object becomes the prediction of adjustment and maladjustment, then the research will necessarily de-emphasize the psychiatric criterion and pursue careful follow-up studies of subjects. Validity will then be determined by real outcomes rather than prediction of outcomes.

Test Construction Problems. Prediction of the psychiatric criterion by instruments which assess the frequencies or intensities with which certain symptoms appear is not enhanced by extensive multiplication of items or the inclusion of filler questions which serve no other purpose than to make the inventory more palatable to the subject. If a single general scale is used for measurement, approximately 40 questions seem to represent a maximum beyond which validity is not increased. Even this may prove much longer than is actually necessary.

Considerable attention has been paid to the problem of item format in the research projects reported. It has been impossible however to reach a decisive conclusion. On an impressionistic basis rather than as deduction from irrefutable evidence, the paired-choice format seems to offer two important advantages, the first of which it shares with the forced-choice type of item.

In both formats it is possible to have random placement of the adjusted and maladjusted choices so that the subject is forced to read the item with some care to determine whether to answer the left or right choice. This makes impossible a careless stereotyped

answering "down the line," which is frequently the case with yes-no items.

A second advantage relates to the problem of item difficulty, and involves a theoretical question in item analysis which has not been treated extensively in the literature. It is of sufficient hypothetical importance to deserve some discussion.

Items are usually selected for these inventories because of their validity in predicting the criterion. A more careful selection procedure would also take into account the intercorrelation of the items. Two items with high validity coefficients may also be very highly intercorrelated. Unless this situation is recognized, the inclusion and scoring of both items may result in giving excessive weight to a single factor. On the other hand, an item which is low in validity may also be highly correlated with the other items of the test. Elimination of such an item because of a poor validity coefficient may result in discarding an item with excellent potentialities as a suppressor variable.

Item difficulty is undoubtedly another significant basis on which item selection should be predicated. By difficulty is here meant the percentage of the total population answering an item in the neurotic or maladjusted direction. Easy items would have few people selecting the maladjusted choice; difficult items would have a large percentage of the population answering in that way.

If an inventory is designed to be effective in connection with the referral of approximately 20% to 30% of the total population, then the most efficient items would probably be those which split the population in approximately the same proportions. Difficulty indices of .20 to .30 would be optimal and would probably result in sharpening prediction within the desired referral range.

This consideration brought the yes-no type of item under some suspicion. This format frequently implies a kind of central cleavage in the continuum represented by the symptom which is the subject of inquiry. Consider this example:

I feel tired and played out most of the time. Yes___ No___

A negative response would imply that the individual was in that part of the continuum which is represented by failure to feel tired at any time up to half way along the path to feeling tired all the time.

Manipulation of this situation is quite simple with the paired-choice format. With this structure, the example item above was made to read:

I feel tired and played out most of the time.	_____	_____ I get tired a little easier now but it never bothers me in my work.

By making the paired statements represent relatively adjacent positions on the scale or continuum, a great many normal individuals with minor symptom difficulties could now answer freely without danger of their being pulled into the neurotic response group and thus build up a higher percentage of false positives. On the basis of actual experience with the above item, it can be reported that it proved much more discriminating in its paired-choice form than in the yes-no form. Quite obviously similar manipulation of any item in paired-choice form can be undertaken to make it more or less difficult.

One of the major weaknesses of the Billet Qualifications Blank in its present form is undoubtedly a matter of item difficulty, a fact which became evident only after item analysis. Those items with the highest validities were items of such low difficulty that only a very small percentage of the total population could be affected by their inclusion in the scoring key. The net result so far as total test scores are concerned was to give the total distribution a very heavy loading of cases in the lower score range. In fact, about 78% of the cases in the Naval Training Center B Study had N-Scale scores of 0, 1, or 2 even though a 39-item key had been employed. This type of score spread is too limited and inflexible to produce the type of discrimination hoped for. In all likelihood, an alteration in the difficulty of the items by changes in structure or content will result in improved prediction.

Other Measurement Techniques. The theoretical weakness implicit in the use of a series of symptom questions as if they represented a single, cohesive scale designed to measure a single factor called maladjustment is quite evident. Maladjustment is manifested in a great variety of ways, and there are not necessarily high degrees of relationship among the surface indications. But it must be recognized that these inventories were designed as screening instruments rather than diagnostic tools. For this reason, without attempting to predict for the whole range of maladjustment, they could be limited quite successfully to a variety of symptoms which are reasonably common to the most frequent diagnostic classifications found in military screening.

Most cases dealt with in military psychiatry, at least in early screening, fall within the general classification of psychoneurosis, with anxiety, hysterical, and neurasthenic reactions predominant. Other types of neurosis, such as compulsive or obsessive trends, or psychotic reactions, or severe character disorders are much less frequent. It has been quite satisfactory, therefore, to employ an inventory in single scale form with primary orientation in the direction

of the hysterical, neurasthenic, and anxiety reactions found most frequently as clinical problems.

This does not mean that a series of diagnostic scales, each of which is specifically directed at some reasonably well defined disability, would not improve prediction or serve a better purpose in the psychiatric evaluation process. The military situation places such high premium on brevity and simplicity that no attempt was made to experiment with instruments of the type represented by the Minnesota Multiphasic Personality Inventory. It was necessary to recognize that these tests would be used in high pressure situations, frequently by classification officers with little psychological training. If the future permits a more leisurely and intensive approach, further experimentation with diagnostic scales is certainly indicated.

It is to be hoped, too, that further research on screening instruments will encompass additional investigation of projective techniques. The one attempt here made to escape from the symptom approach involved the semi-projective technique employed in the Social Judgments Test described earlier. Research results were essentially negative in the sense that prediction of the psychiatric criterion was very inferior to that possible by use of a symptom questionnaire. Both because the technique tried was extremely experimental and contained many weaknesses in its own rationale, and also because significant validation of this type of instrument could only come from use of a different type of criterion, it should not be assumed that projective techniques offer no hope for better prediction. On the contrary, it seems likely that future improvements in test prediction are most apt to come from those procedures which are least dependent on the conscious evaluations and self-ratings of the subject.

Use of Suppressor Variables. Further investigation of the possibilities inherent in the use of a suppression key is indicated. Whatever items or elements are found highly correlated with the predictive test items but uncorrelated with the criterion offer a significant opportunity for improvement in validity. Such items or elements imply that something has entered into the test which does not belong in terms of predicting the criterion. If the extraneous elements can be measured, then they can be subtracted from the test, leaving a residual more purely related to the criterion than before.

Research Methodology. All too frequently, research on the validity of tests emphasizes the degree of success and neglects analysis of the failures. In this respect the present studies are no exception. There is perhaps as much or more to be learned from the study of the unpredicted case as there is to be gained from the study of

cases accurately identified. Who are the false positives, the men who score in the maladjusted direction but are found quite adequate for military service? What are their characteristics? What positive adaptive mechanisms do they possess which permit successful adjustment despite the presence of many significant symptoms? Who are the false negatives, the men who score in the normal direction but prove psychologically unstable and unfit for service? What elements in the test or in the individual prevent detection by the measurement device? These inquiries are of the utmost importance if further test refinement is to be accomplished.

Certainly such inquiries may provide valuable data on whether it is possible to measure the adaptive adjustive elements in personality rather than just the observable weaknesses and deviations. Summation of symptoms present may represent but a fragment of the total picture. What of the assets? What of those defensive adjustments which permit effective adaptation of the organism to the demands of the environment? This affirmative type of measurement may be needed to qualify or amend the results of the more negative type of assessment.

Above all, it seems absolutely essential to carry on future research with a criterion based on actual success in adjustment rather than psychiatric prediction of success. So long as the criterion of psychiatric diagnosis is used, it will be a misnomer to call these inventories tests of adjustment or maladjustment. At least one must postpone calling them that until research has adequately established the validity of the psychiatric prognosis. Major progress will undoubtedly be made when long term studies are undertaken which will proceed as follows: First backward, from the observation of military successes and failures to the specific etiology or characteristics which differentiate such groups; then forward, from the measurement of such differences to the specific validation of the measurements in terms of actual military experience.

PART III

PREDICTION OF SUCCESS IN TRAINING



CHAPTER X

PREDICTION OF SUCCESS IN TRAINING AT PRIMARY OFFICER TRAINING SCHOOLS

General Description of the Studies

As described in Chapter IV, the schools employed in the Navy's wartime primary officer-training program were chiefly indoctrination schools and reserve midshipmen's schools. The purpose of the present chapter is to present evidence on the effectiveness of psychological tests in selecting entrants into these two types of school. In addition, the effectiveness of tests in selecting students for the V-12 colleges will be considered. The V-12 colleges, toward the end of the war, were the principal feeder of the reserve midshipmen's schools. The predictive validity of four aptitude tests and one achievement test is studied in the present chapter.[1]

1. OFFICER QUALIFICATION TEST, FORM 2. This is a brief (one hour) aptitude test devised primarily for the selection of entrants into indoctrination schools. This test was ordinarily administered to civilian applicants for commissions at the offices of naval officer procurement; the brevity of the test was well adapted to the administrative requirements at the procurement offices.

2. ARMY-NAVY COLLEGE QUALIFYING TEST (TEST C-1). This aptitude test is longer than the Officer Qualification Test (two hours), and was devised for the selection of civilians applying for entrance into the V-12 college training program.

3. NROTC SELECTIVE EXAMINATION, FORM C. This is a very brief aptitude test (requiring only 28 minutes of working time), which was in use, prior to the war, for the selection of candidates to the Naval Reserve Officers Training Corps. In the present chapter, a study is reported of the validity of this test for selecting entrants into reserve midshipmen's school.

4. OFFICER CLASSIFICATION TEST, FORM X-1. This two-hour aptitude test was devised primarily for men *graduating* from indoctrination or midshipmen's schools. The four scores obtained from the test—Verbal, Mechanical, Mathematical, and Spatial—facilitated the placement of graduates into particular billets or specialized, advanced schools. Although this test was not devised as a selection measure for entrants into any primary officer-training school, it ap-

[1] This chapter is based upon research reports of NDRC Project N-106 and the College Entrance Examination Board, and upon unpublished studies of the Test and Research Section, Bureau of Naval Personnel (see Appendices C-1 and C-2).

peared desirable, when the opportunity offered, to investigate the predictive value of this test in selecting entrants into reserve midshipmen's school.

5. TEST N-4. This is a comprehensive test designed to measure achievement in the first semester of V-12 college training. Since most of the men succeeding in the V-12 college training program were assigned to reserve midshipmen's school, it seemed desirable to determine to what extent first-semester achievement in V-12 predicts later achievement in reserve midshipmen's school.

The tests mentioned above are described in greater detail in Chapter VII and in subsequent portions of the present chapter.

Although the predictive measures considered in the present chapter are limited almost wholly to tests, it is realized that other devices are also included in any practical selection procedure: physical examinations and interviews generally played an important part in naval personnel selection. In addition, performance during the first few weeks of a school program provides a measure which is frequently useful for predicting subsequent school success; but early school performance should be regarded only as a last resource, since much time, transportation, and expense can be saved if valid prediction is made before training has begun.

The Criteria. Broadly speaking, the success of a selection device can be measured by the success of the selectees. Success, in the studies of the present chapter, refers to training-school success, as indicated by (1) graduation or failure in the training program (a comparatively crude or coarse measure) and (2) by grades in the training-school subjects. It is recognized that the measures of training-school success need supplementation by measures of actual performance in the practical duties for which the school courses provide background or training. The relation between success in school and success on the job is not, in general, sufficiently close to permit a measure of the one to substitute for a measure of the other. Nevertheless, the criterion of success in school has definite practical validity, since a person who fails a training program will not, ordinarily, be permitted to enter the duties for which the program is the official prerequisite.

Samples. The samples studied in the present chapter include 763 officers in an indoctrination school; 427 WAVES officer-candidates in a women's reserve midshipmen's school; 1,342 deck-officer-candidates in a reserve midshipmen's school; and 772 students in seven V-12 schools. An obvious limitation, with reference to generalization of conclusions, is the comparatively small number of schools studied—only one of the thirteen naval indoctrination schools operating

during the war, only one of the seven men's reserve midshipmen's schools, and only seven of the 131 V-12 schools. Limitations of time and size of staff prevented extending the scope of the studies as broadly as might be desired. On the other hand, the number of cases in each study appears large enough to justify some confidence in the results.

The samples used in the studies of the present chapter always involve some selection or restriction, if only because the samples never include applicants who were refused admission to the training schools. Another source of restriction arises when the criterion is school grades (rather than merely graduation vs. failure); because in this case the samples do not include individuals who failed or who (for whatever reason) were dropped from the school. The exclusion of failures and drop-outs is accountable on two grounds: first, quantitative school-grades are generally lacking for failures or drop-outs; and second, the various other data needed for a research study are, in general, much more readily available for regular cases than for failures or drop-outs. The restriction of the samples to entrants or course-survivors causes a decrease in the magnitude of the validity correlations. In most instances, however, it has been possible to estimate a statistical correction for this effect.

Statistical Procedures. The degree of correspondence between the predictive measures and success in training-courses was measured by either the biserial or the product-moment correlation coefficient, usually the latter.

The fact that the samples used in the studies of the present chapter are restricted to entrants or course survivors is unfortunate from the point of view of evaluating a predictive measure. It is desirable to know the correlation between the predictor and the criterion in the total group of *applicants,* since the function of tests, in the selection process, is more largely to determine which of the *applicants* should be accepted or rejected than to predict the precise degree of success of those who are accepted. The correction of the validity correlations for the limited range of ability in the restricted samples was accomplished by the use of Pearson's formula, as presented by Kelley.[2]

Limitations of the Studies. In addition to limitations already mentioned, it may be pointed out that wartime research necessarily tends to place greater emphasis on practical results than on well-rounded comprehensiveness and theoretical depth. A consequent limitation of the studies reported in this chapter is the virtual absence of inquiry into *why* a particular test succeeds as far as it does but no farther. Similarly, all too little investigation has been made of the

[2] Kelley, T. L. *Statistical Method,* pp. 223-225. Macmillan, New York, 1924.

proper emphasis of tests in the total selection process. In the selection of V-12 students, one might ask what is the relative weight that should be given to information from the student's high-school record, from psychological tests, from interviews, and from physical examinations. The studies available at the present time generally leave such comprehensive questions unanswered. An investigation aiming to answer some of these questions regarding the V-12 program is now in progress.

In the following sections consideration will be given first to the prediction of success in indoctrination school, then to the prediction of success in reserve midshipmen's school, and finally to the prediction of success in the V-12 college training program. This sequence corresponds, for the most part, to the chronological sequence in which the separate studies were made, and to the order of peak utilization of the different types of school by the Navy.

Predicting Success in an Indoctrination School by the Officer Qualification Test

The test employed in selecting men for indoctrination school was the Officer Qualification Test. This test is described in Chapter VII. Correlation coefficients were calculated between test scores and grades in each of the indoctrination school courses, Seamanship, Ordnance, and Navigation, and also with a rating known as the Officer-Aptitude Rating. The aptitude rating, based on instructors' judgments, is designed to provide an evaluation of the individual's "officer-like qualities," or the personal qualities deemed desirable in a naval officer. The Officer-Aptitude Rating and the grades in each subject were expressed on a scale from 0 to 4.0, with the passing mark set at 2.5. Finally, correlation coefficients were also calculated between test scores and the student's final average grade. The final average grade is an unweighted average of the aptitude rating and the grades in Seamanship, Ordnance, and Navigation.

The Seamanship course included a great variety of subjects, such as ship organization, naval terminology, ship construction, ground tackle, mooring, communication and orders, inland rules, keeping ship's log, and blinker and flag signals. The Ordnance course included the structure and function of various guns, the fire control problem, and some practice in handling guns. The Navigation course, which included some trigonometry, was intended not to make finished navigators, but simply to insure that as junior officers the men would know something of what was being done during a watch on the bridge.

As mentioned in Chapter VII, the validation results obtained

for the Officer Qualification Test may be regarded in the nature of pre-validation, rather than strictly formal validation. The qualifying circumstances include the following: (1) The time-limit was more liberal than that finally adopted. (This is not considered a serious factor, because even with the shortened time-limit, scores on the Officer Qualification Test by no means depend primarily on speed.) (2) The test was administered to the students *after* they had entered indoctrination school. (This was not considered very important, because the rank-order of a group of men on an aptitude test is not, in general, seriously affected by a uniform educational experience.) (3) The Forms 2 and 3 administered were actually experimental forms containing, in all, 274 items, instead of the 200 eventually retained in the final Forms 2 and 3. The Form 2 and Form 3 scores used in this study represent a re-scoring of the extra-length experimental forms, restricted to the 100 items finally retained in each form. (The interpolation of extra items in the experimental forms is not considered a serious matter, since a special study showed that the item-characteristics [difficulty and item-total correlation] of the retained items in the final forms were practically identical with the characteristics of the same items in the experimental forms.)

RESULTS. Below are listed the correlations between total scores in Forms 1, 2, and 3 of the Officer Qualification Test and final average grades:

Form 1: Correlation with final average grade .51 ($n = 360$)
Form 2: Correlation with final average grade .48 ($n = 403$)
Form 3: Correlation with final average grade .50 ($n = 403$)

The standard deviations of Officer Qualification Test raw scores in the indoctrination school samples are only about .9 as great as in the group of applicants at the offices of naval officer procurement. Statistical correction for this restriction in ability-range [3] would raise the coefficients listed about .04 each. The figures above suggest two conclusions:

1. The Officer Qualification Test is reasonably successful in predicting final average grade in indoctrination school.
2. Forms 1, 2, and 3 of the Officer Qualification Test are substantially equivalent with respect to validity in predicting final average grade in indoctrination school.

Some interest attaches to the correlation between the Officer Qualification Test and the *individual components* of the final average grade. Table 1-x presents the facts for Forms 1, 2, and 3 of the test; the median correlation for the three forms is also given, in the bottom line of the table.

[3] Kelley, T. L., *op. cit.*

It will be observed from Table 1-x that Navigation is the subject for which grades are predicted best (median $r = .50$); Seamanship is next; Ordnance, third; and Officer-Aptitude Rating, last. It is not surprising that the Officer-Aptitude Rating is only very poorly predicted by the Officer Qualification Test, since this rating is based

TABLE 1-x. Correlations between Forms 1, 2, and 3 of the Officer Qualification Test and components of the final average grade (Indoctrination School)

Officer Qualification Test	Correlation with			
	Grade in Seamanship	Grade in Ordnance	Grade in Navigation	Officer-Aptitude Rating
Form 1	.41	.42	.42	.21
Form 2	.45	.34	.51	.07
Form 3	.47	.37	.50	.09
Median r	.45	.37	.50	.09

mainly on personal qualities—which the Officer Qualification Test does not attempt to measure. An additional factor tending toward a low correlation coefficient for the Officer-Aptitude Rating is the low standard deviation of the aptitude-rating distributions (see Table 2-x). Before attempting to interpret further the data of Table 1-x, it will be well to take note of the comparative standard deviations

TABLE 2-x. Means and standard deviations of course-grades and of Officer-Aptitude Ratings (Indoctrination School)

	Class 6		Class 7	
	M	σ	M	σ
Seamanship	3.3	.27	3.4	.30
Ordnance	3.4	.21	3.4	.23
Navigation	3.2	.32	3.1	.31
Final Average Grade	3.3	.18	3.3	.20
Officer-Aptitude Rating	3.2	.16	3.2	.18

for grades in Seamanship, Ordnance, and Navigation. These standard deviations are given (along with the means) in Table 2-x. The standard deviations of grades in Ordnance are appreciably lower than for Seamanship or for Navigation. It is difficult to know whether this is due to a restricted range of ability of students in Ordnance, or to inadequate, non-differential grading. In either

event, however, the low standard deviations for Ordnance grades would tend to result in a lower correlation between these grades and the Officer Qualification Test.

The reliability coefficients of grades should also be considered in interpreting the validity coefficients of Table 1-x. For Seamanship and Ordnance courses, seven or eight weekly grades were obtainable for each man; but for Navigation, only the final average grade was available. Accordingly, the reliability coefficient for Navigation grades could not be calculated. For Seamanship and Ordnance, it was possible to calculate the split-half reliabilities (corrected by the Spearman-Brown formula); these reliability coefficients are as follows:

Reliability of Seamanship grades in Class 6	.86
Reliability of Seamanship grades in Class 7	.92
Reliability of Ordnance grades in Class 6	.80
Reliability of Ordnance grades in Class 7	.75

The lower reliability coefficient for grades in Ordnance partially explains the lower validity of the Officer Qualification Test in predicting grades in Ordnance as compared with Seamanship.

We turn now to the detailed correlations between each subtest of the Officer Qualification Test and indoctrination school grades. In order, however, to avoid an excessively large and detailed table of data, only the median correlation coefficients for the three forms of the Officer Qualification Test are presented, rather than the coefficients for each form separately (see Table 3-x). The use of the median coefficient is justified because the differences in the correlation coefficients for Forms 1, 2, and 3 are quite small and not significant.

TABLE 3-x. Median correlations between subtests of the Officer Qualification Test and school grades (Indoctrination School)

Subtest	Correlation with			
	Grade in Seamanship	Grade in Ordnance	Grade in Navigation	Final Average
Opposites	.37	.27	.34	.36
Mechanical Comprehension	.28	.29	.36	.35
Arithmetical Reasoning	.36	.29	.52	.46

The outstanding facts of Table 3-x are (1) the high correlation coefficient (.52) between Arithmetical Reasoning and grades in Navigation, slightly higher than for the total Officer Qualification Test (see Table 1-x); (2) the low correlations of all subtests with grades

in Ordnance; and (3) the overall superiority of the Arithmetical Reasoning test (correlation with final average grade = .46, as compared with .35 and .36 for Opposites and Mechanical Comprehension).

IMPROVEMENT OF PREDICTION. Although the Officer Qualification Test is, by the evidence, reasonably successful in predicting grades in indoctrination school, still higher correlation coefficients are desirable. It is conceivable that higher coefficients could be obtained through such means as better and more uniform instruction; or through more highly valid grading of students' work; or through a more uniformly high level of motivation among students. Such avenues of improvement, however, are beyond the scope of the present study. It remains to inquire, then, whether higher correlations between the Officer Qualification Test and grades may be obtained by

TABLE 4-X. Simple vs. multiple correlation between parts of Officer Qualification Test and school grades (Indoctrination School)

Officer Qualification Test	Class	Grade in Seamanship		Grade in Ordnance		Grade in Navigation		Final Average Grade	
		r	Multiple R	r	Multiple R	r	Multiple R	r	Multiple R
Form 1	6	.41	.44	.42	.43	.42	.54	.51	.57
Form 2	7	.45	.46	.34	.35	.51	.57	.48	.51
Form 3	7	.47	.47	.37	.39	.50	.57	.50	.55

improvement of the test itself, or by improved use of scores from the separate parts of the test.

The question at once arises whether improved correlations can be obtained by *differential weighting* of the subtests. The maximum improvement which could be expected from altering the weights of the three parts can be ascertained by computing the multiple correlation of the three part-scores with the criteria, and comparing these multiple correlations with the simple, or original, correlations. Table 4-x presents such comparisons.

Table 4-x may be interpreted several ways. If the ideal multiple-regression weights were determined and applied, it appears that grades in Seamanship and Ordnance would not be predicted appreciably better; but grades in Navigation would be predicted more accurately (the correlation coefficient rising from a median r of .50 to a median R of .57); and the final average grade would also be predicted more accurately (the correlation coefficient rising from a median r of .50 to a median R of .55) It may be questioned whether

such gains are worth the extra trouble in applying the special weights (or approximations) required by the multiple regression equation. The weights are not, of course, identical for predicting grades in Navigation as for predicting final average grade. Since the Arithmetical Reasoning subtest is the most valid of the three parts of the test (Table 3-x), it might be better from a practical point of view merely to increase the number of items in this test (thus increasing its relative weight). If this were done, it would be necessary to shorten one or both of the remaining subtests in order to remain within the one-hour limit fixed for the Officer Qualification Test.

The Mechanical Comprehension subtest is only moderately valid for predicting indoctrination grades (Table 3-x); but the lack of a high reliability coefficient for this subtest suggests that increasing the number of items in Mechanical Comprehension might improve its validity perceptibly. Taking the reliability coefficient of the later forms of the Mechanical Comprehension subtest as .75, it is estimated [4] that doubling the length of the Mechanical Comprehension test would raise its correlation with grades in Seamanship from .28 to .30, with grades in Ordnance from .29 to .31, with grades in Navigation from .36 to .38, and with final average grade from .35 to .37. Evidently, merely increasing the length of the Mechanical Comprehension test will not appreciably improve its validity.

Probably the validity of predicting both indoctrination school grades and Officer-Aptitude Ratings could best be improved by the addition of a test or questionnaire having a low correlation (about .05) with the Officer Qualification Test and a definitely positive correlation (about .25) with the criterion. The addition of such a test would raise the correlation between the Officer Qualification Test and final average grade from about .50 to .55.[5] A personality test or questionnaire relating to personal and biographical background might best be able to fulfill the statistical requirements for the hypothetical validity-rise from .50 to .55. If weighted scores on the three subtests of the Officer Qualification Test are employed, the median correlation between Officer Qualification Test and final average grade is itself .55 (Table 4-x); addition of a successful personality or background measure such as described above, would raise this validity-correlation to .59.

The application of multiple-regression weights (or approximations) is somewhat troublesome; and the possibilities described in the previous paragraph would definitely require the use of such

[4] Guilford, J. P. *Psychometric Methods*, p. 422, formula 202. McGraw-Hill, New York, 1936.

[5] Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*, p. 31, formula 321. World Book Co., Yonkers, N. Y., 1932.

weights. Nevertheless, the possibility of a rise in validity from .50 to .59 is sufficiently attractive to justify serious attention to whatever clerical or machine-scoring procedures seem to stand in the way of such a rise.

Predicting Success in a Reserve Midshipmen's School (Women's Reserve) by the Officer Qualification Test

PREDICTIVE MEASURES AND CRITERIA. Forms 2 and 3 [6] of the Officer Qualification Test were administered to three companies of WAVES officer-candidates at a reserve midshipmen's school (women's reserve). The present section will consider evidence on the validity of the Officer Qualification Test, as determined by correlations between the test and course-grades.

The courses taken by the WAVES officer-candidates included Basic Indoctrination and either Communications or Advanced Indoctrination. Each of these courses embodied four subjects, as shown in the following outline:

Basic Indoctrination (symbolized by "I"):
- I-1 (a) Organization
  (b) Naval Law
  (c) Naval Correspondence
- I-2 Personnel
- I-3 History and Strategy
- I-4 Ships and Aircraft

Communications (symbolized by "C"):
- C-1 Naval Communications (General)
- C-2 Fundamentals of Radio
- C-3 Codes and Ciphers
- C-4 Touch Typewriting

Advanced Indoctrination (symbolized by "A"):
- A-1 (a) Organization
  (b) Communications
- A-2 Personnel
- A-3 (a) History and Strategy
  (b) Correspondence and Filing
- A-4 Ships and Aircraft

Of the three companies included in the present study, Companies 1 and 2 took Basic Indoctrination and Communications. Company

[6] These were experimental forms of the Officer Qualification Test, described in Chapter VII.

6 took Basic and Advanced Indoctrination. The numbers of officer-candidates tested in each of these three companies were, respectively, 152, 160, and 115.

RESULTS. Table 5-x gives the correlations between Forms 2 and 3 of the Officer Qualification Test and students' final average grades. The median of the correlations is under .40, which is notably lower than the median of .50 observed in the sample of indoctrination school students.

For Company 1, the final average grade was determined by the formula, $\frac{I + 2C}{3}$ (where "I" is final grade in Basic Indoctrination, and "C" is final grade in Communications); for Company 2, the final average grade was determined by the formula, $\frac{I + 3C}{4}$; for

TABLE 5-X. Correlations between Forms 2 and 3 of Officer Qualification Test and final average grade (Women's Reserve Midshipmen's School)

Officer Qualification Test	Correlation with Final Average Grade	
	Companies 1 and 2[1] (N = 312)	Company 6[2] (N = 115)
Form 2	.35	.39
Form 3	.37	.47

[1] These two companies took Basic Indoctrination and Communications; each correlation coefficient listed is the arithmetical average of the r's for the two companies. The means for the two companies in final average grade were 3.21 and 3.23, respectively; the standard deviations were .20 and .21, respectively. The mean and σ of raw scores on the Officer Qualification Test (for Companies 1, 2, and 6 combined) were 55.5 and 10.9, respectively.

[2] Company 6 took Basic and Advanced Indoctrination. The mean of final average grades for Company 6 was 3.26, the σ was .22.

Company 6, the final average grade was determined simply by averaging final grades in Basic and Advanced Indoctrination. All grades were expressed on a scale from 0 to 4.0, with the passing mark set at 2.5.

It is of interest to note the correlations between the Officer Qualification Test and the individual components of the final average grade. Table 6-x presents these figures, for Forms 2 and 3 of the Officer Qualification Test; the average correlation coefficient for these two forms is also given.

Table 6-x shows that Communications is the subject with which the Officer Qualification Test has the lowest correlation (average r = .29). The reason for this low correlation coefficient does not lie in an exceptionally low standard deviation of grades in Communica-

tions (see Table 7-x). The low correlation is in general accord with other findings of low relationship between scores on general aptitude tests such as the Officer Qualification Test, and grades for the learning of such subjects as radio code and typing.

TABLE 6-X. Correlations between Forms 2 and 3 of Officer Qualification Test and components of final average grade (Women's Reserve Midshipmen's School)

Officer Qualification Test	Correlation with Grades in		
	Basic Indoctrination (N = 427)[1]	Communications (N = 312)[2]	Advanced Indoctrination (N = 115)[3]
Form 2	.40	.28	.34
Form 3	.43	.30	.41
Average r	.42	.29	.38

[1] Average of correlation coefficients for Companies 1, 2, and 6. The use of an average r is justifiable, since the r's for the different companies do not differ significantly from each other.

[2] Average of correlation coefficients for Companies 1 and 2.

[3] Based on Company 6 only.

Since the reliability coefficients of grades in the different courses are pertinent to interpretation of the correlations in Table 6-x, an attempt has been made to estimate these reliabilities. Because weekly grades were not available, the technique of correlating the sum of

TABLE 7-X. Means and standard deviations of grades in Basic Indoctrination, Communications, and Advanced Indoctrination (Women's Reserve Midshipmen's School)

Course	M	σ
Basic Indoctrination [1]	3.3	.25
Communications [2]	3.2	.21
Advanced Indoctrination [3]	3.3	.21

[1] Average of values for Companies 1, 2, and 6.

[2] Average of values for Companies 1 and 2.

[3] Values for Company 6 only.

odd-week grades with the sum of even-week grades could not be applied. Fortunately, grades for each of the four subjects within a course were obtainable; accordingly, the sum of grades in two subjects within a course was correlated with the sum of grades in the other two subjects, and the Spearman-Brown formula was then ap-

plied. For Basic Indoctrination, the sum of grades in I-1 and I-3 was correlated with the sum of grades in I-2 and I-4. For Communications, the sum of grades in C-1 and C-3 was correlated with the sum of grades in C-2 and C-4. For Advanced Indoctrination, the sum of grades in A-1 and A-3 was correlated with the sum of grades in A-2 and A-4. The resulting reliability coefficients [7] are as follows:

Reliability of Basic Indoctrination grades in Co. 1	.82
Reliability of Basic Indoctrination grades in Co. 2	.87
Reliability of Basic Indoctrination grades in Co. 6	.87
Reliability of Communications grades in Co. 1	.63
Reliability of Communications grades in Co. 2	.69
Reliability of Advanced Indoctrination grades in Co. 6	.70

The differences among the reliability coefficients are not sufficient to account for the differences among the validity coefficients of Table 6-x, especially since the lowest reliability coefficients (for Communications) probably represent a serious underestimate of the true reliability.

TABLE 8-X. Average correlations between subtests of Officer Qualification Test and course grades (Women's Reserve Midshipmen's School)

Subtest	Correlation with Grades in: Basic Indoc. (N = 427)[1]	Communications (N = 312)[2]	Adv. Indoc. (N = 115)[3]	Final Average (N = 427)[1]
Opposites	.34	.18	.32	.29
Mechanical Comprehension	.25	.21	.10	.20
Arithmetical Reasoning	.34	.28	.36	.34

[1] Average of correlation coefficients for Forms 2 and 3, in Companies 1, 2, and 6. The use of average r is justifiable, since neither the r's for the different forms nor the different companies exhibit significant differences.

[2] Average of correlation coefficients for Forms 2 and 3, in Companies 1 and 2.

[3] Average of correlation coefficients for Forms 2 and 3, in Company 6.

We turn next to the detailed correlations between each subtest of the Officer Qualification Test and course grades. In order, however, to avoid an excessively large and detailed table of data, only the *average* correlations for the two test forms, and for the various companies, will be presented (Table 8-x).

[7] It is obvious that the basic assumptions for the use of a split-half correlation and for application of the Spearman-Brown formula are only very roughly fulfilled here. In particular, the subjects matched in the odd and even halves of Communications, Naval Communications (General), and Codes-Ciphers in the odd half against Fundamentals of Radio and Touch Typewriting in the even half, fail to fulfill the requirements of similarity and comparability. This results in a spuriously low reliability coefficient.

The two outstanding facts in Table 8-x are (1) the comparatively low correlation coefficients yielded by the Mechanical Comprehension subtest and (2) the comparatively low correlation coefficients for grades in Communications (see also Table 6-x).

IMPROVEMENT OF PREDICTION. Since the validity coefficients of the Officer Qualification Test for WAVES officer candidates are not sufficiently high to be considered satisfactory, some possible means of improvement may be mentioned. (1) A small but possibly worthwhile improvement in validity (roughly .05) could be gained by

TABLE 9-X. Simple vs. multiple correlation between Officer Qualification Test and course grades (Women's Reserve Midshipmen's School)

Officer Qualification Test	Company	Correlation with: Basic Indoctrination Grades r	Basic Indoctrination Grades Multiple R	Communications Grades r	Communications Grades Multiple R	Advanced Indoctrination Grades r	Advanced Indoctrination Grades Multiple R	Final Average Grades r	Final Average Grades Multiple R
Form 2	1	.40	.40	.24	.34			.33	.38
Form 2	2	.41	.42	.31	.34			.37	.38
Form 2	6	.39	.49			.34	.39	.39	.46
Form 3	1	.42	.43	.24	.36			.34	.40
Form 3	2	.41	.45	.36	.38			.40	.42
Form 3	6	.47	.55			.41	.46	.47	.53
Average Correlation		.42	.46	.29	.36	.38	.42	.38	.43

applying multiple-regression weights to the individual subtests of the Officer Qualification Test (Table 9-x). (2) Another source of improvement might be the substitution of a clerical ability test in place of the Mechanical Comprehension subtest. (Eliminating the Mechanical Comprehension subtest would not reduce the validity of the Officer Qualification Test in the sample of WAVES.[8]) Finally (3) the addition of a test relating to personality or biographical background or both might also be advantageous.

[8] The following figures support this statement: the multiple correlation between final average grade on the one hand, and Opposites, Arithmetical Reasoning, and Mechanical Comprehension on the other, is .43; the multiple correlation, omitting Mechanical Comprehension, is practically the same, .42. Similar multiple correlation coefficients for the separate courses taken by the WAVES students are as follows: for Basic Indoctrination, .46 compared with .45; for Communications, .36 compared with .32; and for Advanced Indoctrination the correlation was .42 for both cases.

Predicting Success in a Naval Reserve Midshipmen's School (Deck)

The study to be reported below is based on 1,342 deck-officer-candidates in a class at a naval reserve midshipmen's school. This class convened on July 5, 1944. All but a few of the candidates had qualified for reserve midshipmen's school through training in the Navy's V-12 college program.

PREDICTIVE MEASURES. The purpose of the study is to determine the validity of various measures available for selecting entrants to reserve midshipmen's school. The predictive measures employed are as follows:

1. Officer Qualification Test, Form 2.
2. Officer Classification Test, Form X-1.
3. NROTC Selective Examination, Form C.
4. Test N-4.
5. Average grade during Indoctrination (a three-week period preceding the two regular terms of reserve midshipmen's school).

Of the five measures listed above, the first two have already been described (Chapter VII). The NROTC (Naval Reserve Officer Training Corps) Selective Examination, Form C, is composed of four tests. The first test, Multiplication, consists of one hundred simple multiplication problems (two-place numbers multiplied by one-place numbers); four minutes are allowed for this test. The second test is Block Counting; again, four minutes are allowed. The third test is the familiar vocabulary test (a key word is given, and the examinee selects from five alternatives the one which means the same as the key); time allowance, five minutes. The fourth and last test is Arithmetical Reasoning; time, fifteen minutes. Test N-4 is a comprehensive, five-hour general achievement test administered to V-12 students at the beginning of their second semester. The indoctrination grade is the average of grades for the three-week indoctrination period in Navigation, Ordnance and Gunnery, Seamanship, and Damage Control. Grades are expressed on the usual Navy 0-4.0 scale. Each subject was weighted 2, except Damage Control, which was weighted 1.

CRITERIA. As criteria by which to test the validity of the predictive measures described above, use was made, first, of the final academic average; and second, of graduation vs. failure.

The final academic average in reserve midshipmen's school is based on academic grades during the three periods into which the complete course is divided: the indoctrination period (three weeks), the first term of reserve midshipmen's school (six weeks), and the second term of reserve midshipmen's school (six weeks). The average of grades during the indoctrination period is combined with grades during the first week of the regular term to form the first "weekly

grade." At the end of each of the six-week terms, a term-grade is determined in each subject, by giving the average of weekly grades a weight of 2, and the final examination of the term a weight of 1. The final grade in each subject is obtained by averaging the two term-grades. The final academic average is obtained by averaging the subject-grades for the two six-week terms, giving each subject a weight as follows: Navigation, 2; Ordnance and Gunnery, 2; Seamanship and Communications, 2; Damage Control, 1; and Recognition Training, 1.

The graduation vs. failure criterion distinguishes merely between those who passed, and those who failed for academic reasons. Of the total academic failures, about 80 per cent are dropped at the end of the first three-week (indoctrination) period. At this time the men are advanced from apprentice seamen to the status of midshipmen. In general, if a candidate had received a failing mark in more than one of the indoctrination-period subjects, he was retained only if his aptitude test scores indicated a reasonably high level of ability and if there appeared to be extenuating circumstances, such as extensive hospitalization, to explain his unsatisfactory work. Ratings of officer-aptitude were also taken into account. College transcripts, containing V-12 records, were consulted in special cases, but they were not used regularly because standards among institutions were found to vary, and the records lacked uniformity. No use was made of C-test or N-test scores in deciding whether to retain a student.

The Officer-Aptitude Ratings mentioned above are ratings primarily of the various personal qualities (as distinguished from intellectual abilities) deemed important for naval officers. The Officer-Aptitude Rating has not been included as a criterion in the present study for several reasons: (1) the available internal evidence is not reassuring as to the validity of the ratings: 80 per cent of the ratings fall within 3.2 and 3.4, inclusive, and the standard deviation of the distribution is quite small (.13); (2) the predictive measures in the present study are designed primarily as measures of ability, not personality; and (3) the correlation between grades and the school's over-all evaluation of the candidate (the final multiple grade) is .97, so that very little is lost by excluding the Officer-Aptitude Ratings.

The comparative weight of course grades, aptitude test scores, and Officer-Aptitude Ratings in deciding whether a candidate was to be passed or considered an academic failure has not been studied. No doubt course grades played a major role; but aptitude test scores and Officer-Aptitude Ratings were also taken into account, especially for doubtful cases. Because of the school authorities' familiarity with and confidence in the NROTC Selective Examination, this test was given greater weight than other aptitude measures in determining

whether a candidate was to be retained in the reserve midshipmen's school or separated. To the extent that aptitude test scores were used to determine retention or separation, the correlation between pass-fail and aptitude test scores may be considered spuriously high. From the point of view of validating the aptitude tests, it would have been better had the test scores not been consulted when the decision was being made to pass or fail a candidate.

RESULTS. The principal findings of the study of reserve midshipmen are given in Table 10-x. This table shows the correlations be-

TABLE 10-X. Correlations between selection measures and final academic average (Reserve Midshipmen's School [Deck])

Selection-Measure	Correlation with Final Academic Average[1]
NROTC Selective Examination, Form C	.46
Officer Qualification Test, Form 2	.45
Officer Classification Test: Verbal + Math.	.52
Officer Classification Test: Verbal	.35
Officer Classification Test: Mechanical	.27
Officer Classification Test: Mathematical	.49
Officer Classification Test: Spatial	.29
Test N-4 (V-12 Comprehensive Objective Test)	.54
Average Grade for Indoctrination Period	.80

[1] All correlation coefficients are based on 1,109 cases, except the coefficient for Test N-4, which is based on 639 cases. The reasons for this exception are, first, that the N-test record was not always available at the reserve midshipmen's school; and second, that those entering the reserve midshipmen's school directly from the Fleet had never taken Test N-4.

tween the various selection measures and the criterion of final academic average. The correlation coefficients in this table are based on the 1,109 cases graduating in a single class of the reserve midshipmen's school.

Table 10-x may repay some scrutiny. The highest correlation coefficient is between final academic average and average grade for the initial three-week indoctrination period. This correlation is .80, a very serviceable degree of relationship.[9] It would be highly desir-

[9] Two technical considerations affect the correlation of .80. First, it will be recalled that indoctrination grades provided a principal basis for eliminating academic failures from the midshipmen's school. This implies that the remaining sample was less variable in indoctrination grade than it was before, and such a reduction in variability tends to reduce the correlation between indoctrination grades and final academic averages. On the other hand, it will be recalled that the average of grades during the indoctrination period was the principal component of the first "weekly grade" of the regular term. This means that the indoctrination grade entered into the final academic average, so that the correlation of .80 represents a small amount of self-correlation. (The nominal weight of indoctrination grades in the final academic average is 1/24.) No precise determination of the net effect of the two counterbalancing influences seems called for.

able, however, to predict success in reserve midshipmen's school at a time earlier than the end of the indoctrination period. The next highest correlation coefficient in Table 10-x is for Test N-4 (.54). Test N-4, normally coming early in the student's second semester of the V-12 college training program, provides prompter information than indoctrination grades, but this is at the cost of a drop in the correlation coefficient from .80 to .54. The Officer Classification Test, being an aptitude test, can be administered before the student has begun his V-12 college work.[10] The Mathematical score on this test yields a validity correlation of .49, the Verbal score, a correlation of .35, and the simple sum of Mathematical and Verbal scores,[11] a correlation of .52. This last correlation coefficient is practically as high as that for Test N-4. Beside the possibility of early administration, a further advantage of the Officer Classification Test is that the Mathematical and Verbal sections of this test together require only 65 minutes of working time, whereas the N-test requires 5 hours.

It is possible that selected portions of the N-test, like the selected parts of the Officer Classification Test, might also yield a satisfactory correlation with the final academic average. Unfortunately, the relation of part-scores of the N-test to the final academic average was not investigated. In any case, the Officer Classification Test would still have the advantage of being suitable for earlier administration.

In the present sample the Officer Classification Test was administered to students at the beginning of their stay in the reserve midshipmen's school; whereas the N-test was administered on November 30, 1943 (seven months earlier). It is doubtful that such a comparatively brief time-interval would (for an aptitude measure such as the Officer Classification Test) appreciably affect the correlation with final average grade. The possibility of such an effect, however, might well be investigated.

The correlation of .46 for the NROTC Selective Examination with final academic average is interesting, since the NROTC test is so very brief, requiring only 28 minutes of working time. In another study of the NROTC test, similarly favorable results were obtained: a correlation of .49 with grades in a deck officers' school ($N = 216$), and a correlation of .51 with grades in an engineering officers' school ($N = 193$). But there is no particular advantage in

[10] A separate test, the Army-Navy College Qualifying Test (or C-test), was developed for admission to the V-12 program (see Chapter VII and page 202 of this chapter); but none of the students in the reserve midshipmen's class of the present study had taken the C-test, having entered college before the first administration of the test.

[11] Scores on the parts of the Officer Classification Test are expressed on a standard score scale, in which the mean of an unselected or standard sample of students in indoctrination or reserve midshipmen's schools equals 50, and the standard deviation equals 10.

extreme brevity when time is available for a test which can give more reliable results or a wider variety of useful subtest scores.

To what extent can the results in Table 10-x be improved by combining various test-scores? The following two multiple correlation coefficients indicate the answer to this question:

Let 1 = final academic average
2 = score on NROTC Selective Examination, Form C
3 = score on Verbal + Mathematical sections of the Officer Classification Test
4 = average grade for the indoctrination period
Then $R_{1.234} = .816$ (the simple correlation, r_{14}, equals .80)
$R_{1.23} = .543$ (the simple correlation, r_{13}, equals .52)

These multiple correlation coefficients discourage the hope that the combining of scores from different tests will yield sufficient increase in validity to justify the extra time and expense.

The general conclusion from the findings set forth above would seem to be that if an early prognosis of performance in reserve midshipmen's school is desired, the Mathematical and Verbal sections of the Officer Classification Test are the most serviceable.

The recommendation of the Officer Classification Test for the prognosis of scholastic achievement in reserve midshipmen's school raises the question of the relation of this test to the C-test (used in selecting candidates for the V-12 program). The salient fact in this connection is that the C-test was used in a situation where the rejection rate was high; applicants for the V-12 college training program were numerous, and the average level of mental ability was below the standard deemed suitable for the program. The C-test, if only for practical reasons, needed to be considerably easier than the Officer Classification Test. There was no point, however, in accepting students for the V-12 program if they could not succeed in later training; and this later training took place, most commonly, at the various reserve midshipmen's schools. From this point of view, the Officer Classification Test might have proved useful as a means of checking up on borderline or doubtful cases. For such purposes, the Officer Classification Test (or, at least, the Mathematical and Verbal parts of this test) could have been administered at the offices of naval officer procurement at the time that the V-12 applicant came up for physical examination and interview. It is necessary to remember that the present study is based on deck-officer-candidates in a reserve midshipmen's school. The Officer Classification Test might prove less successful as a selective or verifying measure for students who plan to go from V-12 training to a supply corps school, or to a professional school for chaplains, dentists, or doctors.

We turn now to an examination of results according to the *second* criterion, that of graduation vs. failure. Tables 11-x through 14-x show the distributions of test scores for those who graduated, and for those who failed (for academic reasons) in the reserve midshipmen's school. In each case the test scores for the academic failures are on the average considerably lower than for the graduates; but the

TABLE 11-x. Scores on NROTC Selective Examination (Form C), for graduates and for academic failures (Reserve Midshipmen's School [Deck])

Score	Number of Graduates	Number of Academic Failures
220-229	2	
210-219	2	
200-209	1	
190-199	8	
180-189	30	
170-179	58	2
160-169	91	
150-159	157	4
140-149	176	12
130-139	186	20
120-129	147	23
110-119	113	37
100-109	76	37
90-99	34	27
80-89	14	30
70-79	7	17
60-69	5	13
50-59	2	6
40-49		1
30-39		3
20-29		1
N	1,109	233
M	137.5	103.4
σ	24.4	26.1

overlap of the distributions emphasizes the desirability of pushing validity coefficients beyond the level reached by even the best of current tests.

It is generally recognized that shortcomings in a criterion may adversely affect the correlation between a test and the criterion. An elaborate study of the validity of grades in the reserve midshipmen's

school of the present study could not be undertaken. At least with respect to stability from one semester to the next, however, the grades in the reserve midshipmen's school appear satisfactory: correlations between first- and second-semester grades in Navigation, Seamanship and Communications, Ordnance and Gunnery, and Damage Control range from .67 (for Navigation) to .77 (for Ordnance and Gunnery). These correlations are based on a sample of 364 cases, randomly selected from the graduates of a single class.

TABLE 12-x. Scores on Officer Qualification Test, for graduates and for academic failures (Reserve Midshipmen's School [Deck])

Score[1]	Number of Graduates	Number of Academic Failures
74-76	1	
71-73	2	
68-70	9	
65-67	12	
62-64	18	
59-61	36	1
56-58	59	2
53-55	85	4
50-52	139	12
47-49	141	10
44-46	189	27
41-43	170	35
38-40	150	49
35-37	57	38
32-34	31	38
29-31	8	14
26-28	1	3
23-25	1	
N	1,109	233
M	46.5	39.6
σ	7.56	6.12

[1] The scores are expressed on a standard score scale, in which the mean score of an unselected or standard sample of applicants at the offices of naval officer procurement equals 50, and the standard deviation equals 10.

As a means of summarizing the data of Tables 11-x through 14-x, Table 15-x is presented. Table 15-x gives the biserial correlation (r_{bis}) between test-scores and the criterion of graduation vs. failure. The biserial correlation coefficients range from .30 (for the Mechanical section of the Officer Classification Test) to .68 (for the NROTC

TABLE 13-X. Scores on separate sections of the Officer Classification Test, for graduates and for academic failures (Reserve Midshipmen's School [Deck])

	Verbal Section		Mechanical Section		Mathematical Section		Spatial Section	
Score [1]	No. of Graduates	No. of Academic Failures	No. of Graduates	No. of Academic Failures	No. of Graduates	No. of Academic Failures	No. of Graduates	No. of Academic Failures
77-79			1		3			
74-76			8		4		5	
71-73	5		6	1	18		7	
68-70	17		6	1	22	1	27	
65-67	22		27	3	34	1	44	1
62-64	28	3	32	2	88	1	92	4
59-61	33		63	4	62	1	61	5
56-58	57	3	64	7	83	5	146	12
53-55	76	3	88	7	125	12	109	18
50-52	76	5	136	23	120	13	180	26
47-49	128	24	97	15	154	14	104	18
44-46	139	15	113	22	126	24	134	33
41-43	114	30	165	27	92	27	55	26
38-40	194	38	102	31	77	41	60	32
35-37	159	65	82	29	59	42	31	12
32-34	48	36	61	24	25	31	28	22
29-31	13	10	26	21	14	13	12	11
26-28		1	18	9	2	5	7	8
23-25			13	6	1	1	2	3
20-22			1	1		1	3	1
17-19							1	1
N	1,109	233	1,109	233	1,109	233	1,108	233
M	45.4	39.4	46.5	41.1	50.3	40.6	51.3	43.2
σ	8.92	6.42	9.96	9.52	9.50	7.52	9.46	9.19

[1] The scores are expressed on a standard score scale—see Table 12-X.

Selective Examination, Form C). For convenience in comparison, the product-moment correlation coefficients of Table 10-x, based exclusively on those who graduated, are included in Table 15-x. The high value of r_{bis} for the NROTC test is probably due, as has been previously explained, to a comparatively heavy weighting of scores

TABLE 14-x. Raw scores on Test N-4 (V-12 Comprehensive Objective Test), for graduates and for failures (Reserve Midshipmen's School [Deck])

Raw Score	Number of Graduates	Number of Academic Failures
675-699	1	
650-674		
625-649		
600-624	1	
575-599	2	
550-574	1	
525-549	7	
500-524	7	
475-499	16	
450-474	17	
425-449	37	
400-424	37	1
375-399	54	9
350-374	73	5
325-349	100	12
300-324	100	17
275-299	68	19
250-274	61	22
225-249	33	31
200-224	19	21
175-199	5	18
150-174		5
N	639	160
M	341.1	264.0
σ	74.6	59.0

from this test when deciding whether to separate or to retain a student at the end of the three-week indoctrination period. The reason for the comparatively high value of r_{bis} for the Spatial section of the Officer Classification Test is not known. A possible hypothesis is that a curvilinear relation exists between ability on this test and scholastic succcess in reserve midshipmen's school—the hypothesis

being, further, that beyond a certain minimum spatial ability, the advantage of additional ability is small; whereas below a certain level, the disadvantage is significant. Such a situation would lead to a higher value of the biserial correlation for the total group (including both failures and graduates), than of the product-moment correlation for the group of graduates only. This hypothesis, of course, requires experimental verification.

RECOMMENDATIONS. The main recommendations from the present study have already been indicated. Of the various measures investigated, the three yielding the most serviceable prediction of final average grade in reserve midshipmen's school are (1) average grade during the three-week indoctrination period, (2) score on the N-test

TABLE 15-X. Biserial correlation between test scores and the criterion of graduation vs. failure—together with product-moment correlations between test scores and final academic average (Reserve Midshipmen's School [Deck])

Test	Biserial r with Criterion of Graduation vs. Failure ($N = 1{,}342$)	Product-Moment r with Final Academic Average ($N = 1{,}109$)[1]
NROTC Selective Examination, Form C	.68	.46
Officer Qualification Test, Form 2	.50	.45
Officer Classification Test: Verbal + Math.	.56	.52
Officer Classification Test: Verbal	.38	.35
Officer Classification Test: Mechanical	.30	.27
Officer Classification Test: Mathematical	.55	.49
Officer Classification Test: Spatial	.46	.29
Test N-4 (V-12 Comprehensive Objective Test)	.57	.54

[1] See note to TABLE 10-X.

(administered early in the second semester of the V-12 program), and (3) a score based on the Mathematical and Verbal sections of the Officer Classification Test. The results from multiple correlation discourage the hope that combining scores from different tests will yield sufficient increase in validity to justify the extra labor and expense. For an early prediction of final academic average in reserve midshipmen's school, the Officer Classification Test is the most useful. The average grade during the three-week indoctrination period yields the most accurate prediction, but indoctrination grades are not available until the candidate is on the very threshold of entrance into reserve midshipmen's school. It is quite possible that a test or questionnaire relating to the candidate's personality or biographical

background would prove a useful addition to the aptitude and ability measures studied in this section.

Predicting Success in the V-12 College Training Program

The Navy V-12 college training program may be described as a program at the pre-officer level. The program provided appropriate college training to qualified high school graduates who, upon successful completion of four or more semesters of college work, were to be trained as officers in the U. S. Naval Reserve. The first input of men into the V-12 program took place on July 1, 1943. This first input included (1) men who were in the earlier V-1, V-5, and V-7 college programs, in inactive reserve status; (2) Army Enlisted Reserve Corps enrollees who expressed a preference for naval service; (3) enlisted men entering the program from active duty in the Navy; and (4) civilian high school graduates. Enlisted men in the Navy were admitted to the V-12 program on recommendation of their commanding officers; among the requirements for recommendation were an examination for physical fitness, graduation from high school, and a score of at least 85 on the old (O'Rourke) Navy General Classification Test or 110 on the Marine Corps General Classification Test (after the new Navy General Classification Test became available, a score of at least 58 was required on this test). The selection process for civilian high school graduates included a number of steps, as follows: (1) the Army-Navy College Qualifying Test (Test C-1); (2) a physical examination; (3) an evaluation of the candidate's previous high school (or college) record; (4) interviews by two officers at an office of naval officer procurement; and finally, (5) an evaluation of all the collected information about the candidate by a committee composed of an educator, a naval officer (a senior officer at the office of naval officer procurement), and a "good common-sense civilian."

The primary purpose of the present section is to consider the validity of the Army-Navy College Qualifying Test (Test C-1) for predicting scholastic success in the first semester of V-12 training. Attention will also be given to a comparison of performance of high school graduates from civilian vs. enlisted status. All data are limited to men entering in the "first increment" (July 1, 1943). It should be emphasized, at the outset, that the study of the C-1 test and the V-12 college training program is still in progress, so that the present account can be no more than preliminary. But a limited number of questions will be answered, at least tentatively; and a more complete presentation will be available later.

Predictive Measure and Criteria. Test C-1, the college qualifying test, is a two-hour examination divided into four sections, which

may be designated as Verbal, Science, Reading, and Mathematical. Examinees were advised to spend 30 minutes on the first section (Verbal), and 30, 25, and 35 minutes, respectively, on the other three sections. For a more detailed description of Test C-1, see Chapter VII.

The criteria against which the validity of the C-test has been studied include: (1) scores on a comprehensive objective test, Test N-4, administered a few weeks after the end of the first semester of V-12 training; and (2) the first-semester average grade. Both the N-4 test scores and average grades were expressed on a scale running from 0 through 10.

The first-semester average grade is based on grades in the following subjects: Mathematics (algebra and trigonometry), Physics, English, History, Engineering Drawing, and Naval Organization. Although no directive on the matter was issued by the Bureau of Naval Personnel, most colleges appear to have weighted the various grades according to the number of "contact hours" for each subject —namely, 5, 4, 3, 2, 2, and 1, respectively.

The N-test is a five-hour examination divided into English (100 minutes), Physics (50 minutes), Mathematics (90 minutes), and History (60 minutes). The sections on Physics, Mathematics, and History were designed as broad general-achievement tests in these subjects. Since they were administered in all the V-12 colleges (131 in number), these tests could not cover the special features of emphasis in any local curriculum. The test in English included (1) a subsection on language usage (25 minutes); (2) a subsection on the detection of superfluous terms and phrases in an exposition (20 minutes); (3) a subsection on paragraph reading (25 minutes); unlike the paragraphs in the C-1 test, most of these paragraphs are moderately technical; and (4) a subsection containing the same kinds of item as the Verbal section of the C-1 test (30 minutes).

It may be observed that the paragraph-reading and last-named subsections of the N-test are similar to two sections of the C-test. On this ground it may be objected that these two parts of the English test should not be included in a battery employed as a criterion for the C-test. In deference to this objection, and because it is not convenient to isolate the score on the two overlapping parts of the English test from the score on the other two parts, the entire English score will be eliminated from the total N-4 score in certain statistical presentations.

SAMPLE. The sample used to investigate the validity of Test C-1 includes students at seven V-12 institutions: two liberal arts colleges (one in New England, one in the South), three universities (one in the Middle West, two in what might be termed the "upper South"),

one Midwestern teachers college, and one technical institute. The total number of cases studied is 772. Men with incomplete records were eliminated from the sample, as were also pre-medical and pre-dental students, who followed a first-term curriculum somewhat different from that outlined above. Of the total 772 cases, 637 came into the V-12 program from civilian life; the remaining 135 were enlisted men who came into the program from the Fleet. Since the Fleet cases (for practical reasons) were not given the C-1 test, it was convenient (in this preliminary investigation) to limit most of the statistical treatment to the principal sample of 637 cases. These 637 cases will be referred to as the "total civilian sample."

Validity of the C-test

CORRELATION BETWEEN C-TEST AND N-TEST. The correlation between the total scores on Test C-1 (the college qualifying test) and Test N-4 (the comprehensive achievement test), in the total civilian sample of 637 cases, is .70. This correlation was also calculated separately for the college containing the largest number of V-12 students of civilian origin (N = 204); here the correlation between scores on the total C-test and the total N-test was found to be .68.

The two correlation coefficients, .70 and .68, may be considered somewhat inflated because of the similarity of two parts of the English section of the N-test with the Verbal and Reading sections of the C-test. It has therefore seemed advisable to calculate the correlation between the C-test and the N-test *exclusive of English.*[12] These correlations are .64 and .61, respectively. These corrected coefficients are probably somewhat too low, since the elimination of the *total* English section from the N-test, instead of only the two overlapping parts of the English section, represents an over-elimination.

The correlations of .64 and .61 apply to the sample of cases who were not only admitted to the V-12 college training program, but were retained at least a full semester in that program. The primary purpose of the C-test, however, is not so much to predict the performance of students who are admitted to the V-12 program, as to discriminate between qualified and unqualified applicants. From this point of view, it is reasonable to inquire what the correlation

[12] The formula employed is: $\frac{r_{cx} + r_{cy} + r_{cz}}{\sqrt{3 + 2(r_{xy} + r_{xz} + r_{yz})}}$ where c represents the total C-score, and x, y, and z represent the three parts of the N-test, exclusive of English. This formula involves the assumption of equal standard deviations for x, y, and z. For the sample of 637 cases the standard deviations are, respectively, 2.1, 2.2, and 2.1; and for the sample of 204, 1.9, 2.2, and 2.1. Since no highly precise determination is required, these standard deviations may be regarded as virtually equal.

between C-score and N-score would be in a sample including all applicants. The standard deviation of C-scores of the sample of 637 cases is 13.0[13]; of the sample of 204 cases, is 13.1; and of all the applicants for the Navy V-12 program (123,206 cases) is 23.1. Correcting the correlations of .64 and .61 to the values that would prevail in a sample with a C-score standard deviation of 23.1, we obtain[14] the values of .83 and .80, which are high enough to indi-

TABLE 16-X. Detailed correlations between Test C-1 and Test N-4[1] (V-12 College Training Program)

N-Test / C-Test	N-4 Math.	N-4 Physics	N-4 English	N-4 History	N-4 Total
C-1 Math.	.53 *.57*	.47 *.52*	.23 *.31*	.24 *.22*	.49 *.58*
C-1 Science	.32 *.38*	.60 *.60*	.19 *.19*	.09 *.15*	.42 *.44*
C-1 Verbal	.19 *.17*	.16 *.16*	.68 *.64*	.44 *.44*	.51 *.47*
C-1 Reading	.17 *.18*	.16 *.17*	.28 *.27*	.34 *.34*	.31 *.31*
C-1 Total	.42 *.46*	.49 *.52*	.62 *.62*	.44 *.46*	.68 *.70*

[1] The italic figures in the lower-right corner of each cell are for the total sample of 637 cases entering V-12 from civilian life; the figures in the upper-left corner are for a single-college group of 204 cases.

cate a very serviceable degree of relation between C-test (administered before admission to the college training program) and N-test (administered after a full semester of college work).

Table 16-x presents detailed correlations between the individual sections of the C-test and the separate sections of the N-test, both for the total civilian sample of 637, and the single-college sample of 204. None of these correlation coefficients has been corrected for

[13] By way of comparison, a standard deviation of 13.0 on Test C-1 is equivalent to a standard deviation of 73 on the College Entrance Examination Board's Verbal Scholastic Aptitude Test. On this test, the standard deviation of a typical unselected group of applicants seeking admission to colleges making use of the College Entrance Examination Board's tests is 100.

[14] Kelley, T. L. *Statistical Method*, p. 225, formula 186. Macmillan, New York, 1924.

restricted range.[15] Especially interesting are the correlations in the columns headed "N-4 Math." and "N-4 Physics." In the total civilian sample, the C-1 Math. section predicts N-4 Math. better than does the total C-1 test (.57 vs. .46); and the C-1 Science section predicts N-4 Physics better than does the total C-1 test (.60 vs. .52). Similar results may be observed in the single-college sample of 204 cases.

A noteworthy fact is the failure of the Reading subtest, despite its superior "face validity," to yield as high a validity-correlation with N-4 History as the Verbal part of the C-test. Perhaps the memory-factor involved in History achievement is more closely related to the Verbal than to the Reading section of the C-test. This

TABLE 17-X. Comparison of multiple and simple correlations between individual parts of Test N-4 and sections of Test C-1 (V-12 College Training Program)

	Sample of 637 Cases		Sample of 204 Cases	
	Multiple R	Simple r	Multiple R	Simple r
N-4 Math. vs. (C-1 Math., C-1 Science)	.60		.55	
N-4 Math. vs. C-1 Math.		.57		.53
N-4 Math. vs. C-1 Total		.46		.42
N-4 Physics vs. (C-1 Science, C-1 Math.)	.68		.66	
N-4 Physics vs. C-1 Science		.60		.60
N-4 Physics vs. C-1 Total		.52		.49
N-4 History vs. (C-1 Verbal, C-1 Read.)	.48		.47	
N-4 History vs. C-1 Verbal		.44		.44
N-4 History vs. C-1 Total		.46		.44

suggests that a reading test which emphasizes delayed memory may be desirable.

It has seemed worth while to inquire to what extent the combining of separate sections of the C-test, by the technique of multiple correlation, may still further improve the prediction of scores on separate parts of the N-4 achievement test. Table 17-x presents the findings.

From Table 17-x it appears that a worth-while improvement in prediction of N-4 Physics (from .60 to .68, or from .60 to .66) can be gained by the multiple correlation technique; the

[15] Such a correction would raise all the correlation coefficients, tending to raise the higher coefficients in Table 16-x more than the lower. The *differences* among the corrected correlation coefficients would be at least as great as among the uncorrected, and it is chiefly on these differences that the interest of the present section centers.

other gains in Table 17-x are less impressive or important. In any event, it is clear that the unweighted sum of scores on the separate sections of the C-test is *not* always the most serviceable means for predicting achievement in each separate subject. This is true especially for Mathematics and Physics. In view of the importance of mathematics and physics in the training of most prospective naval officers, it appears that scores for the separate parts of the C-test might well be taken into account in granting or denying admission to applicants to the V-12 Program.

CORRELATION BETWEEN C-TEST AND AVERAGE COLLEGE GRADE. We turn now to the correlation between C-1 scores and average grades for the first semester of college work. In connection with these data, two points must be borne in mind: (1) It is well known that standards of grading are not uniform from instructor to instructor, from subject to subject, or from school to school. Thus, differences in grading-standards among the seven institutions supplying the sample of 637 civilian cases are likely to reduce the coefficients for this group to a level below that in the single-college group of 204 cases. (2) The range of ability in the group retained in the V-12 program through the first term is much narrower than in the sample applying for admission; and as mentioned previously, the purpose of the C-test is principally to discriminate between *applicants* who are qualified and those who are not. A correction for restricted range of ability is therefore applicable. In making this correction, the standard deviations of C-scores for the restricted sample were taken from Table 19-x. The standard deviations for the applicants are as given below:

C-1 Total Score	23.1
C-1 Math.	5.5
C-1 Science	7.1
C-1 Verbal	11.9
C-1 Reading	4.3

The first of the figures just listed (23.1) is based on the total sample of 123,206 applicants for the Navy V-12 program; the remaining figures are based on a representative sample of 1,500 cases.

Table 18-x presents the correlation between C-test scores and average grade, both for the sample of 637 cases and the single-college group of 204. Accompanying each original correlation coefficient is the coefficient corrected for restricted range of ability on the C-test.

In Table 18-x the correlation coefficients for the single-college sample of 204 cases are more readily interpretable than those for the sample of 637, since the latter sample is especially likely to

suffer from inter-institutional differences in grading standards. For this reason, the discussion below will be limited to the sample of 204. In this sample, the correlation between total score on the C-1 test and average college grade is .37; corrected, this becomes .58. The correlation coefficient of .37 must be considered rather low; the corrected coefficient of .58 is not higher than is sometimes reported for a college sample without any correction. The correlation coefficients for the separate parts of the C-test are, as might be expected, generally lower than for C-1 total score. An exception occurs, however, in the case of the Mathematical section of C-1, for which the uncorrected correlation with average grade is .38. This is reasonably high for an individual subtest, and is probably explained by the fact that Mathematics and Physics figure very heavily in the average college grade (see page 202). It may be noted that the C-1 Verbal

TABLE 18-X. Correlations between C-1 scores and average grade (V-12 College Training Program)

Test	Sample of 637 Cases		Sample of 204 Cases	
	Raw Correlation with Average Grade	Corrected[1] Correlation	Raw Correlation with Average Grade	Corrected[1] Correlation
C-1 Math.	.29	.41	.38	.55
C-1 Science	.23	.30	.27	.35
C-1 Verbal	.18	.26	.21	.29
C-1 Reading	.14	.26	.20	.37
C-1 Total	.31	.50	.37	.58

[1] Corrected for restricted range of ability in the sample—see text.

and Reading sections yield lower correlation coefficients than either C-1 Science or Mathematics, and that the Reading section displays no special superiority over the Verbal.

Because of the differences among the correlations for individual sections of the C-1 test (see Table 18-x), it was thought that the multiple correlation between average grade and the four sections of C-1 might be appreciably higher than the correlation for C-1 total. Use of the multiple correlation technique raises the raw r of .31 for C-1 Total (in the sample of 637) to .343; and raises the raw r of .37 (in the sample of 204) to .432. The second rise might be worth the time and expense involved in applying multiple regression weights; the former is not. Very likely the most practical recommendation would be simply to increase the length (and thus the relative weight) of the Mathematics and Science sections of the C-test. The present Reading section of the C-test might be eliminated entirely with little

or no apparent effect on the reliability or validity of the total test.[16] If desirable to reduce still further the relative weight of the Verbal section, the score on this section might be divided by an appropriate, convenient constant (such as 2 or 3), before adding the Verbal score to the Mathematics and Science scores of the C-test.

All the validity correlations in this preliminary study are based on men in the initial V-12 input. It will be interesting to observe in the later and more complete investigation whether validity correlations for later increments, when the V-12 program was presumably running more smoothly, are any higher.

SUPPLEMENTARY STATISTICAL DATA. For possible use in the interpretation of the various findings reported above, means and standard deviations of various test and criterion scores are given in Table 19-x. Intercorrelations between the criteria (N-test scores and average grades) are given in Table 20-x.

TABLE 19-x. Means and standard deviations of C-test scores, N-test scores, and average grades (V-12 College Training Program)

Measure	Sample of 637 Cases		Sample of 204 Cases	
	M	σ	M	σ
C-1 Total	93.7	12.96	91.6	13.07
C-1 Math.	18.8	3.69	18.1	3.40
C-1 Science	25.5	5.31	24.0	5.25
C-1 Verbal	34.2	8.08	34.3	8.42
C-1 Reading	15.2	2.29	15.2	2.21
N-4 Total[1]	5.7	1.92	5.6	1.77
N-4 Math.[1]	5.8	2.08	5.5	1.86
N-4 Physics[1]	5.6	2.19	5.1	2.18
N-4 English[1]	5.8	2.04	5.8	1.98
N-4 History[1]	5.9	2.11	6.3	2.06
Average Grade[1]	5.0	1.98	4.8	2.07

[1] Original score or grade converted to a scale from 0 through 10.

INTERPRETATION. The chief contribution of this section is the finding that achievement in mathematics and in physics (as measured by the N-test) can be better predicted by scores on the Mathematics and Science sections of Test C-1 than by the raw score on the total C-1 test. A still better prediction, especially for physics, is possible by use of a weighted combination of the Mathematics and Science sections of the C-test. These facts imply that the Mathematics and

[16] In later forms of the Army-Navy College Qualifying Test the Reading section has been eliminated and the Mathematics section has been increased in length.

Science sections of the C-test might well receive special consideration when admitting, or denying admission to, candidates for the V-12 program. With reference to future forms of the Army-Navy College Qualifying Test, the implication is that the Mathematics and Science portions of the C-test should be allowed to have greater relative weight in determining the total test score. These conclusions are in line with the current tendency in civilian practice to place increased

TABLE 20-X. Correlations between criteria: N-test scores and average grades (V-12 College Training Program)

N-Test	Correlation with Average Grade, in Sample of 637 Cases	Correlation with Average Grade, in Sample of 204 Cases
N-4 Math.	.50	.55
N-4 Physics	.34	.39
N-4 English	.28	.29
N-4 History	.31	.38
N-4 Total	.49	.53

emphasis on aptitude for mathematics and physical science when selecting students for engineering curricula. Since, in the Navy, some officers require less engineering aptitude than others, question arises whether the scoring and content of the C-test should be revised so as to yield two or three different total scores, each best adapted to predict training-success for a major group of officers (e.g., deck, engineering, and supply corps).

Performance of Enlisted Men Who Entered the V-12 Program

As mentioned in the introduction to the V-12 study, some men entered the V-12 program from enlisted status. Since a different selection procedure was applied for enlisted men, it is of interest to compare the V-12 performance of these men with the performance of those entering the V-12 program from civilian life.

In making this comparison, it is desirable to maintain reasonable homogeneity in as many factors as possible. In the present study, it has been possible to maintain a certain degree of homogeneity with respect to the student's home region (northern vs. southern) [17]; and,

[17] By "home region" is meant the geographical area in which the man attended high school. The following states were included as "southern": Virginia, West Virginia, North Carolina, Kentucky, Tennessee, South Carolina, Georgia, Florida, Alabama, Mississippi, Louisiana, Arkansas, Oklahoma, Texas, New Mexico, and Arizona. All other states were included as "northern."

in one comparison, to maintain, also, identity with respect to the V-12 college attended. Table 21-x presents the findings.

From Table 21-x it is clear that the Fleet-origin cases made N-test scores substantially equal to those of the civilian-origin cases; but they made better average college grades. In explaining this discrepancy, two facts appear pertinent. (1) The civilian-origin cases were selected, first of all, on the basis of C-test scores, with a secondary screening based on a rating of the high school record and a rating for officer-aptitude qualities. The Fleet-origin cases, on the other hand, were selected primarily on the basis of job efficiency and favorable personal characteristics (such as industriousness, dependability, fluency, agreeableness, etc.), with a secondary screening on intelligence-test score. Since the C-test is more highly correlated with

TABLE 21-X. Average N-test score and average grade for civilian-origin and fleet-origin cases (V-12 College Training Program)

College Attended	Home Region	Origin	N	Test N-4		Average College Grade	
				M^1	σ^1	M^1	σ^1
X [2]	Northern	Civilian	204	5.6	1.77	4.8	2.07
X	Northern	Fleet	57	5.6	2.40	5.8	2.21
Seven [3]	Northern	Civilian	472	5.8	1.89	5.0	1.80
Seven	Northern	Fleet	107	5.7	2.33	5.9	2.00
Seven	Southern	Civilian	165	5.5	1.98	4.9	2.41
Seven	Southern	Fleet	28	5.1	2.26	5.8	2.11

[1] Original score or grade converted to a scale from 0 through 10.
[2] A New England liberal-arts college.
[3] For a description of the seven institutions included in the present study, see pages 202 and 203.

N-test scores than is job efficiency or favorable personality characteristics, the civilian cases have an initial selection advantage in respect to N-test scores. On the other hand, the primary selection of the Fleet-origin cases on such traits as industriousness, dependability, and agreeableness, probably leads to comparatively higher college grades than to N-test scores. (2) The civilian cases were, in general, closer to their high school days than the Fleet cases. As already mentioned, the N-test could not reflect local curricular peculiarities, but had to restrict itself mainly to the broader, more essential aspects of each subject. As such, the N-test doubtless reflects high school accomplishment to some extent—to an extent greater than do college grades, which hinge more closely on performance in the particular, local curriculum. Thus it seems likely that the

civilian cases enjoyed a relative advantage on the N-test. The greater closeness of civilian-origin cases to high school days is an incidental factor, of no fundamental significance. Similarly, the probably greater personal agreeableness of Fleet-origin cases is also incidental, and of no fundamental significance so far as true academic achievement is concerned. Whether or not the civilian- and Fleet-origin cases are genuinely or basically equal in the abilities tapped by the N-test, and whether the two groups are really different in respect of the academic abilities underlying college grades, depends on the influence of these incidental factors. Unfortunately, the available data do not permit a quantitative appraisal of the various factors involved.

From the data on the Fleet- vs. civilian-origin cases it is tempting to draw some conclusion regarding the comparative effectiveness of the procedures used to select these two groups of cases. But any such comparative evaluation requires consideration of questions relating to sampling and quotas. Specifically: Was the total number of civilians who were qualified for V-12 the same as the total number of Fleet cases who were qualified (the larger the number of qualified individuals, the easier the task of selecting capable men)? How many of the qualified Fleet individuals were unavailable, in a practical sense, because of active combat duty or distance from shore stations? Was recruitment (which converts qualified individuals into *applicants*) equally effective for civilian and Fleet groups? Was recruitment more selective (in the sense of attracting a higher proportion of superior applicants) in one group than the other? Was it necessary to accept lower levels of ability among civilians (in order to meet State quotas) than of Fleet applicants (to whom no State quotas applied)? These are, at best, difficult questions, to which no answer can be attempted in this preliminary report. A possible difference in the academic motivation of Fleet vs. civilian cases and a possible tendency by V-12 instructors to give greater attention to Fleet men are two additional complicating factors.

Summary

The four main groups studied in the present chapter include (1) 763 officers in an indoctrination school; (2) 427 WAVES officer-candidates in a women's reserve midshipmen's school; (3) 1,342 deck-officer-candidates in a reserve midshipmen's school; and (4) 772 students in seven V-12 schools. A brief summary of findings is given below.

INDOCTRINATION SCHOOL. 1. The Officer Qualification Test is reasonably successful in predicting grades in the indoctrination

school. This is indicated by a correlation coefficient of about .50 between total test score and final average grades.

2. Of the three parts of the Officer Qualification Test, the Arithmetical Reasoning subtest is the most successful. Scores on Arithmetical Reasoning correlate slightly higher with grades in Navigation (.52) than do scores on the total Officer Qualification Test (.50). In addition, scores on Arithmetical Reasoning are more highly correlated with final average grade (.46) than scores on either of the other two parts of the total test (.36 and .35, respectively).

3. The validity of the Officer Qualification Test could be improved somewhat (about .05) by the use of multiple regression weights in combining scores from the separate subtests. It is suggested that the addition of a test or questionnaire relating to personality or biographical background might also lead to improved prediction of grades in indoctrination school.

WOMEN'S RESERVE MIDSHIPMEN'S SCHOOL. 1. The Officer Qualification Test is only moderately successful in predicting final average grades in a women's reserve midshipmen's school (correlations with final average grade are slightly under .40). Of the three subtests of the Officer Qualification Test, the Arithmetical Reasoning and Opposites subtests predict final average grades more successfully than the Mechanical Comprehension subtest (correlations of .34, .29, and .20, respectively). Grades in the course on Communications are predicted less well by the Officer Qualification Test (correlation about .29) than are grades in the other courses (correlations about .40); this is in conformity with other findings of a low relationship between general aptitude tests and such functions as the learning of radio code and typewriting.

2. The validity of the Officer Qualification Test could be improved somewhat (about .05) by use of multiple regression weights in combining scores from the separate subtests. It is suggested that validity could also be improved by substituting a clerical ability test for the Mechanical Comprehension test, in the form of the Officer Qualification Test administered to WAVES officer-candidates. The addition of a test or questionnaire relating to personality or biographical background might also prove advantageous.

RESERVE MIDSHIPMEN'S SCHOOL. 1. The final academic average in reserve midshipmen's school (deck) can be predicted best from grades during the three-weeks indoctrination period preceding the two regular terms of the school (correlation = .80). For those entering reserve midshipmen's school from the V-12 college training program, an earlier prediction can be made from scores on the N-test (normally administered shortly after the first semester of the V-12 program); this test yields a validity coefficient of .54. A still

earlier prediction is possible by application, to all prospective reserve midshipmen, of the Officer Classification Test. Scores on the Mathematical section of this test correlate .49 with final academic average, and the simple sum of standard scores on the Mathematical and Verbal sections of the test yields a validity coefficient of .52. Scores on the NROTC Selective Examination (Form C) and the Officer Qualification Test correlate with final academic average to the extent of about .45.

2. Combining scores from different tests (such as the Officer Classification Test and the NROTC Selective Examination) fails to lead to validity-gains of practical importance.

V-12 College Training Program. 1. The V-12 study includes a total of 772 students in attendance at seven colleges. The predictive measure investigated is the Army-Navy College Qualifying Test (Test C-1). The validity criteria for Test C-1 are Test N-4 (a comprehensive objective test administered a few weeks after the end of the first semester of the V-12 program), and average college grades during the first semester of V-12. Because part of the English section of Test N-4 is quite similar to the Verbal part of C-1, the English section has generally been excluded from Test N-4 in determining validity coefficients for Test C-1.

2. In a sample of 637 cases, the correlation between Test C-1 and Test N-4 (minus the English section) is .64; in a group of 204 cases attending a single college (these cases are included in the 637), the correlation is .61. Correction of these coefficients for the restricted range of C-scores in the sample (as compared with applicants for V-12 training) yields values of .83 and .80, respectively. Correlation coefficients such as those cited in this paragraph are high enough to justify the use of Test C-1 in selecting applicants for the V-12 college program.

3. The Mathematical section of the C-test, by itself, correlates higher with N-test scores in Mathematics than does the total C-test (the correlations are about .55 and .44, respectively). Similarly, the Science score of the C-test, by itself, correlates higher with N-test scores in Physics than does the total C-test (the correlations are about .60 and .50, respectively.[18] It is clear that the total C-test score, obtained by simply adding scores on the separate sections of the test, is by no means the best available measure for predicting achievement in separate subjects.

4. The use of multiple regression weights in combining scores

[18] The four correlation coefficients cited are all "original" coefficients, uncorrected for restricted range of ability in the sample. Correction for restricted range would raise all the coefficients, tending to raise the higher coefficients more than the lower, and thus to increase the indicated differences.

from the separate parts of the C-test raises the correlation with Physics to about .67 (an increase of .07), but fails to yield gains of practical importance for the other parts of the N-test.

5. The correlation between total C-test scores and average first-semester grades is rather low, .37 in the sample of 204 cases. Corrected for the restricted range of C-test scores in the sample (as compared with applicants for V-12 training), this becomes .58.

6. The Reading section of the C-test tends to yield a lower validity correlation than the Verbal section. There seems to be no special advantage in retaining Reading in the C-test in its present form.

7. The use of multiple regression weights in combining scores for the separate parts of the C-test raises the correlation with average grades to a limited extent (between .03 and .06).

8. A sample of 135 men admitted to V-12 from enlisted status was compared with the sample of 637 admitted from civilian life. The enlisted sample made average scores on the N-test equal to those of the other sample, but higher average college grades. Some possible explanations for this discrepancy are considered.

9. It is suggested that the C-test might advantageously be supplemented by the Mathematical and Verbal parts of the Officer Classification Test, for the purpose of checking up on borderline or doubtful cases.

10. Modification of the C-test should be in the direction of giving increased weight to the Mathematics and Science sections of the test. A Reading subtest placing emphasis on delayed memory might prove more useful than the present Reading section. Consideration should be given to a possible revision of the scoring and content of the C-test so as to yield two or three different total scores, each best adapted to predict training success for a major group of officers (e.g., deck, engineering, and supply corps).

11. It is emphasized that study of the validity of selection procedures in the V-12 college training program is still in progress, so that the section of this chapter devoted to the V-12 program is preliminary in nature.

The various selection tests yield validity correlations between (approximately) .40 and .60, without correction for restricted range in the samples studied. Such correlations are high enough to indicate a useful degree of efficiency; at the same time, such correlations are low enough to indicate that either additional tests or other selection devices are needed to raise the efficiency of prediction to a still higher level. It is suggested that a test or questionnaire relating to personality or biographical background may constitute a useful addition.

The test of Mechanical Comprehension in the Officer Qualifica-

tion Test was found to be of slight value for use with WAVES officer-candidates; the substitution of a clerical ability test may be recommended. The Reading section of Test C-1 proved, in general, slightly inferior to the traditional verbal type of material (such as opposites, analogies, and double definitions). There is no apparent advantage in retaining the Reading test in its present form. The arithmetical reasoning or mathematical aptitude type of test proved especially useful in predicting performance both in indoctrination school, in reserve midshipmen's school, and in V-12 school. The value of taking account of scores on individual tests (as distinguished from total or overall scores only) was made especially clear in the study of V-12 students.

Looking to the future, one may hope, especially, for an extension of criteria beyond course grades to actual performance of practical billet duties; to a more comprehensive program of investigation, which will determine the proper place of each selection device or procedure in relation to the others; and to a more ambitious research program aiming at deeper and broader theoretical understanding, without which progress remains at a merely technological level. Also needed are simple clear experimental demonstrations, to provide evidence of the superiority of scientific personnel procedures over "common sense" and over practices having no superior sanction than use in the past.

CHAPTER XI

PREDICTION OF SUCCESS IN ADVANCED OFFICER TRAINING PROGRAMS

As indicated in Chapter IV, it became necessary to supplement basic instruction of naval officers with various types of specialized training. Instruction given in primary schools was designed to provide a general orientation in naval customs and discipline and to acquaint officers with the fundamentals of navigation, seamanship, ordnance, and similar subjects which are presumably a part of the general background of all line officers. Beyond that, advanced training was necessary to prepare certain officers for specialized billets such as tactical radar, radar specialist, air combat intelligence, or diesel engineering, and for duty aboard special types of ships such as submarines, motor torpedo boats, and various amphibious craft.

Selection requirements for each training program were developed by the Officer Selection Unit in the Bureau of Naval Personnel, and an interviewing program was established in the primary schools to funnel officers into programs for which they were best qualified. Similar interviewing programs were set up at several of the operational training schools where officers were further distributed to specific billets aboard a given type of ship. At the request of the Officer Selection Unit, the Test and Research Section conducted research on the prediction of success in advanced officer training programs. Most of this research was "after the fact" in the sense that the Test and Research Section was called in to determine the validity of existing procedures rather than to assist in initiating the classification program, to study unselected groups, or to carry out a complete validation study. In this chapter the nature of the validation problem, the approach used, the principal findings, and some of the implications growing out of the research completed during World War II will be described.

Nature of the Problem

PURPOSE OF RESEARCH. The research was carried out primarily to obtain answers to two broad types of questions: (1) What qualifications and attributes are most highly related to success in various advanced officer training programs? (2) What are the minimum qualifications and attributes required for satisfactory achievement in such training programs? It was expected that the answers to these

questions would provide the basis for modification of selection requirements for the advanced training programs.

POPULATION. The samples available for study were intact classes in advanced officer schools, classes which had been organized following varying amounts of selection on *a priori* grounds. Officers in these classes had come principally from two sources: (1) reserve midshipmen's schools—mostly young men in their early twenties with two to four years of college training and little or no civilian work experience, and (2) indoctrination schools—mostly men commissioned directly from civilian life with no previous naval experience. In addition, there were a few officers from the Naval Academy and a number of officers assigned to advanced schools following a tour of duty, either ashore or afloat. Officers from the Naval Academy were eliminated from most of the studies, however, or treated in separate analyses.

It should be noted that the samples used in these studies were all preselected. Scholastic aptitude, education, and civilian experience had, of course, been considered in the original selection of officer candidates. In addition, officers were distributed to the various advanced schools on the basis of interviewers' recommendations which were in turn based on aptitude test scores, age, education, civilian occupational experience, record of achievement in basic schools, and the duty preference of the candidate (see Chapter II). Consequently, samples of officers in any given advanced school sometimes differed markedly from the basic officer population in mean or variance or both, on the prediction variables and in the interrelationships between various prediction variables.

The data available on qualifications of officers in advanced schools was characteristically spotty. No uniform qualifications blank existed to which research workers could turn for complete and comparable information on each officer, regardless of the advanced school studied. The Officer Qualifications Record Jacket had not been in operation sufficiently long at the time these studies were made to be of significant value. Duty Recommendation Forms and Training School Reports, developed by the Officer Selection Unit, proved to be of value in the later studies, although even these forms were not available in significant numbers until after most of the studies had been completed. Consequently, data had to be obtained from nonstandardized school records, which contained whatever background data the officers in charge of the particular school in question deemed important, or by means of special questionnaires prepared for the research study.

NATURE OF ADVANCED TRAINING PROGRAMS. Limitations of time and personnel made it impossible to carry on research for all ad-

vanced training programs. An attempt was therefore made to study those given highest priority by the Officer Selection Unit on the basis of numbers of officers involved and complexity of the prediction problem. The two general types of training programs following primary training schools are described in Chapter IV.

Advanced Training for Specific Types of Duty. Schools were established to provide training for specialized types of duty such as tactical radar, radar specialist, air combat intelligence, or fire control. Following such training, an officer might be assigned to any one of several types of duty stations. For example, a large number of officers were sent to communications training, after which they were sent to destroyers, amphibious commands, staff duty, advanced bases, and various other types of duty stations. Training in this type of advanced schools consisted largely of classroom instruction, although attempts were made to utilize laboratories, shops, and mock-ups simulating operating conditions.

Operational Training for Duty on Specific Types of Ships. Operational and precommissioning training courses were developed to enroll officers directly out of basic training, or in some cases from advanced schools of the type described above, and to train them for duty on a specific type of ship, such as a submarine, motor torpedo boat, or landing ship, medium (LSM). In some cases such training was geared toward a particular billet aboard the specific type of ship, as commanding officer or executive officer of an amphibious craft, or assistant gunnery officer, communications assistant, or assistant navigator on a destroyer. For certain highly specialized craft, particularly submarines and motor torpedo boats, a general training for all-round duty on that type of ship was given. Operational training usually involved a relatively short period of classroom orientation and instruction, often given in a brief introductory portion of the total training course, while considerable emphasis was placed on the practical application of knowledges and skills in operational training aboard ship.

Differentiation within Advanced Training and Operational Training. Within each of these two broad types of advanced training programs, a further differentiation can be made on the basis of the degree of technical knowledge (knowledge of mathematics, physics, engineering) required for successful completion of the training course. The curricula of all the schools include certain elements of a technical nature, and strictly speaking, the schools should be placed on a continuum in this respect since they vary in degree of technicality. It may be helpful, however, to distinguish between highly technical and semi-technical training programs in each of the two main categories. Hence, the types of training programs may be

outlined as follows, with the specific programs studied by the Test and Research Section listed under broadly descriptive headings.

1. Specialized training schools for specific types of duty:
   a. Highly technical—pre-radar, engineering, advanced fire control, advanced ordnance and gunnery.
   b. Semi-technical—tactical radar, fighter director, air navigation, air combat intelligence, communications.
2. Operational training for duty on specific types of ships:
   a. Highly technical—submarines, destroyers.
   b. Semi-technical—attack transports, minesweepers, landing ships, medium, motor torpedo boats.

Criterion Measures of Success in Training. The criteria used in these studies will be evaluated in later sections of this chapter; here they will simply be described. In general, three different kinds of end-of-course estimates of achievement were used as criteria: (1) training course grades assigned by instructors, (2) ratings by supervising officers, and (3) achievement test scores. In some instances more than one type of criterion was used in a single study. Quite frequently rank in class was the index of success actually used, but this in turn was based on one or more of the above-mentioned measures. For the most part, the criteria used in these studies were simply the criteria used by the schools themselves in evaluating the achievement of their officer students. Little effort was made by the representatives of the Test and Research Section, except in the case of the studies at Naval Training School (Tactical Radar), to develop criteria other than those used by the schools themselves. Each of the three types of criteria referred to above is described briefly in turn.

Course Grades. Navy training programs customarily assigned daily or weekly grades on each aspect of the instructional program. These grades were usually based on officers' performance on nonstandardized, usually non-objective, tests (prepared and administered by instructors and covering the course content assigned for the immediately preceding period). Sometimes the grades included marks on notebooks or other exhibits prepared in accordance with instructors' directions; and sometimes they were based on instructors' ratings of performance on equipment, as in the combat information center mock-up, or the attack teacher at the submarine school. Grades assigned in specific courses were averaged, with varying systems of weighting particular courses, to obtain the "academic average." This average, sometimes in combination with officer aptitude

ratings, constituted the final grade which determined an officer's standing in his class. In the Test and Research Section studies, the composite academic grade was usually used, although in some cases grades in specific courses were used as partial criteria.

Ratings by Supervising Officers. Ratings by supervising officers were characteristically used for two purposes:

1. to assess those qualities of leadership, work-habits, and military bearing generally subsumed under the heading of "officer aptitude", and
2. to evaluate quality of performance afloat (both under instruction and during shakedown).

Relatively simple rating forms were generally provided, patterned somewhat after the Officer Fitness Report; and supervising officers were required to indicate whether a given officer was unsatisfactory, average, good, or outstanding on the points in question. Seldom was the rating analytical, designed to provide internal checks, or accompanied by more than the simplest effort to train raters in the use of the rating instruments. The raters usually had reasonably good opportunity to observe the performance of the officers they rated, but they were not provided with adequate standards of evaluation nor with satisfactory instructions for systematic sampling of students' performance.

Achievement Test Scores. Only in the tactical radar and pre-radar studies were standardized objective final achievement examinations developed. Scores on such examinations were used as independent criteria and were also combined with other grades and ratings in the determination of final average grade.

Methods of Analysis and Their Limitations

The analyses were primarily correlational; various types of coefficients of correlation were utilized to estimate the relationship between predictors and criteria of success. In some cases analysis of variance was used where the predictive categories were not on a continuum, as in the case of major field in college, or civilian occupational or age groups. Predictors were correlated (1) with success in school defined on an all-or-none basis as "pass" or "fail" and (2) with success as defined in terms of degree of achievement of the passing students, disregarding failing students or placing them in the lowest category of success.

It is important in interpreting the results presented in the next portion of this chapter to be cognizant of the numerous limitations of these analyses. One of the principal outcomes of this phase of the research program may be that of sensitizing those responsible for

research to the obstacles which must be overcome before results meriting a high degree of confidence can reasonably be expected. Consequently, the following limitations, mentioned incidentally in preceding paragraphs, are listed and described below.

PRESELECTION OF TRAINEES. As in almost all validation research conducted by the Test and Research Section, trainees whose qualifications and performance were studied had already been selected, to varying, and largely unknown, degrees, on certain of the variables whose predictive value was the object of study. Never was a random sample or a sample representative of the entire basic officer school population sent to a school as part of an experimental procedure to determine what qualifications were most highly related to success in that training program. Candidates were selected on the basis of multiple criteria, with interviewers applying varying standards to meet immediate quota demands. In addition, certain factors not under investigation, such as degree of motivation of the officer being classified, were allowed to operate in the selection of the samples, but the extent of their influence was not measured or controlled. Such preselection complicates greatly the problem of interpreting results so as to infer what relationships between predictive variables and criteria would hold for an unselected population.

LIMITED NATURE OF PREDICTIVE DATA. Certain predictive variables can scarcely be said to have been thoroughly evaluated; data on civilian occupational experience, for example, was very scanty (limited only to the general title of an officer's civilian occupation, as engineer, foreman, salesman, or production inspector, with no indication of the specific type of work done, number of employees supervised, salary, or any other index of level of skill or degree of responsibility). Findings regarding the predictive value of civilian occupational experience must therefore be considered as merely indicative of what might be expected had more refined measures been used.

Furthermore, a number of plausible hypotheses were never submitted to experimental test. Pressure for immediate results and limitations in numbers of competent personnel available for research work prevented study of many promising leads. For example, no measures of interest were used systematically and evaluated scientifically. Officers in basic schools were asked to express their preference for duty. But these expressions were biased by suggestions from interviewing officers, based partially on their knowledge of billet quotas at the time the advice was given. Evaluation of officers by their fellow officers, self-evaluations, and use of biographical inventories are other possible predictors not subjected to thorough investigation.

INADEQUACY OF CRITERIA. As indicated above, the criterion used was ordinarily the composite of grades and ratings, or the class rank derived therefrom, used by the school as its index of achievement. These indices were by definition the sole measures of success as long as attention was directed exclusively to selection for existing schools. Nevertheless, they are subject to the following criticisms as criteria on the basis of which to generalize concerning the predictive value of various qualifications data.

1. No systematic study was made of the validity of the objectives and course content of the schools, in terms of their relationship to the actual operational job to be performed. A preliminary study was begun in which success in school was to be checked against success in duty at sea, but the war ended before the validation study could be carried out.

2. Aside from the question of validity of the schools' objectives and content, the validity of grades and ratings on specific aspects of many of the training programs was open to serious question. Grades and ratings were largely subjective; instructors' biases and prejudices, as well as chance error, undoubtedly entered into the evaluations in varying degrees. Moreover, they were not known to be comparable from school to school nor from class to class in a given school—making inter-class comparisons and the pooling of data from several classes precarious.

3. The problem of prediction was further complicated by the schools' use of composite criteria of success, often composed of disparate elements. Officers responsible for training were not always in agreement concerning the relative weight to be given the so-called practical factors as compared to theoretical factors in evaluating officers' technical competence. Nor did they agree regarding the weight to be given to technical competence, however determined, as against officer aptitude or officer-like qualities. Consideration was seldom given to the effect of dispersion on the nominal weights employed, so that functional weights might, and often did, vary markedly from those intended.

4. Finally, in some instances criterion data were contaminated; predictive data were available to the raters at the time the criterion ratings were made. In the destroyer study the screening board's rating was used despite this fact because the board was known to have attempted to minimize the degree of contamination and its judgments were felt to be highly superior in other respects to the alternative criteria available.

It should be pointed out that in some instances an attempt was made to evaluate and improve grades and ratings before using them as criteria; but frequently lack of time to conduct preliminary analyses and unwillingness of school officers to question their evaluation techniques thwarted such efforts. In the early stages of its work, the Test and Research Section itself was probably unaware of the serious limitations of the criteria commonly in use; but as the evidence accumulated, the officers engaged in research came to recog-

nize that the establishment of satisfactory criteria was the major problem, to be given precedence over further refinement of predictors, more elaborate statistical techniques, and other means of improving the validation studies.

Lest the foregoing account leave the impression that the studies were valueless, it should be pointed out that not all these criticisms apply to every advanced training program and that efforts were made to select predictors and criteria at any given school so as to minimize these limitations. Some proposed studies, such as the LST (landing ship, tank) and armed guard programs, were dropped because of the practical impossibility of securing reasonably adequate criteria. At the same time, it must be admitted that only in the tactical radar and pre-radar studies was a real effort made to develop an acceptable criterion; and in that fact lies the greatest weakness of the studies on the prediction of success in advanced training programs for officers.

Findings

Correlation coefficients [1] between success in training and certain aptitude tests are shown in Table 1-XI. Means and standard deviations for these tests are given in Table 2-XI. For most of the schools, the criterion of success in training was final school grades. In a few cases, indicated by footnotes to the table, performance ratings were used.

Prediction of Success in Training for Specific Types of Duty

HIGHLY TECHNICAL SCHOOLS. Although it is difficult to generalize on the basis of data from studies which are not strictly comparable, the following statements seem warranted from an analysis of validation studies in five advanced officer schools with highly technical curricula. (Table 1-XI, Part I.)

1. Success in advanced officer schools with highly technical curricula can be predicted reasonably well by use of the Mechanical and Mathematical Sections of the Officer Classification Test. (In order to interpret this statement properly, it should be noted that, except in the case of the air navigation and pre-radar programs, no predictive variables other than scores on the Officer Classification Test were investigated.)

2. There is some evidence that for such a highly technical course as pre-radar, a test of achievement in a specialized field of knowledge (in this case a General Physics Test) may surpass the Officer Classification Test in effectiveness of prediction. The obtained correlation coefficient between scores on the General Physics Test and success in pre-radar train-

[1] Analyses of a non-correlational character are not presented in tabular form.

TABLE 1-XI. Correlation coefficients between scores on aptitude tests and success in training in fourteen officer training programs. With a few exceptions (indicated by footnotes) the criterion was final school grades

Type of Training Program	N	Officer Classification Test: Part I Verbal	Part II Mech.	Part III Math.	Part IV Spatial	Total	CIC Aptitude Total	General Physics	General Math.
I. Highly Technical Schools									
1. Pre-Radar [1]	74/120	.09	.37	.37	.27			.54	.48
2. Advanced Fire Control	128	.23	.45	.31	.26				
3. Diesel Engineering	64	.27	.51	.43	.28				
4. Air Navigation	64	.17	.13	.53	.20				
5. Ordnance and Gunnery	205	.35	.34	.41	.30				
II. Semi-technical Schools									
1. Tactical Radar [2]	83/306	.32	.19	.49	.08		.54		
2. Fighter Director	90						.50		
3. Air Combat Intelligence	58	.18	.02	.42	.52				
4. Aviation Engineering	124	.33	.33	.20	.19				
5. Communications	124	.17	.26	.33	.17				
III. Operational Training Programs									
1. Submarine School	293								
a. r with final class standing		.27	.22	.34	.24	.40			
b. r_{bis} with pass-fail criterion		.26	.23	.54	.37	.47			
2. Destroyer [3]	335	.30	.30	.36	.38	.46			
3. Motor Torpedo Boat	108	.01	.26	.36	.14	.31			
4. LSM Course Grades	95	.32	.29	.24	.11	.29			
LSM Afloat Ratings [4]	95	.22	.20	.04	.05	.13			

[1] Statistics for the parts of the Officer Classification Test were obtained from a sample of 74 cases. Statistics for the General Physics and General Mathematics Tests were obtained from a different sample of 120 cases.

[2] Scores for the parts of the Officer Classification Test were available for only 83 officers who had been assigned to NTSch (Tactical Radar) from Midshipmen and Indoctrination Schools. Scores on the CIC Aptitude Test were obtained on all men in 3 classes, N = 306.

[3] Criterion consisted of ratings assigned by a "Screening Board."

[4] Ratings assigned by afloat training instructors.

TABLE 2-XI. Means and standard deviations of scores on aptitude tests for officers enrolled in fourteen officer training programs

Type of Training Program	N	Officer Classification Test[1]								CIC Aptitude Test[2]		Pre-Radar Tests[2]			
		Part I Verbal		Part II Mechanical		Part III Mathematical		Part IV Spatial		Total		General Physics		General Mathematics	
		M	σ	M	σ	M	σ	M	σ	M	σ	M	σ	M	σ
I. Highly Technical Schools												51.8	7.9	29.8	5.7
1. Pre-Radar[3]	74/120	54.1	8.8	58.4	11.6	65.1	9.1	59.0	8.1						
2. Advanced Fire Control	128	53.3	10.5	56.5	10.3	54.0	11.3	52.4	11.1						
3. Diesel Engineering	64	48.4	8.7	61.5	9.4	59.5	9.9	56.7	9.2						
4. Air Navigation	64	47.3	10.9	53.4	8.5	49.1	8.8	53.9	7.9						
5. Ordnance and Gunnery	205	54.0	10.0	54.7	10.8	53.5	12.7	52.1	10.2						
II. Semi-technical Schools															
1. Tactical Radar[4]	83/306	59.2	7.63	61.9	9.8	66.5	8.1	62.1	7.7	95.4	20.3				
2. Fighter Director	90									84.1	18.3				
3. Air Combat Intelligence	58	60.1	8.9	43.2	9.6	45.2	8.7	45.2	8.7						
4. Aviation Engineering	124	49.1	8.2	64.3	9.1	59.3	11.1	59.2	8.6						
5. Communications	124	54.1	10.2	50.8	11.0	48.6	11.0	50.3	10.1						
III. Operational Training Programs															
1. Submarine School	293	51.5	8.1	58.7	10.3	59.4	10.1	56.6	8.4						
2. Destroyer	335	53.4	9.9	52.5	10.0	53.8	10.5	51.7	9.7						
3. Motor Torpedo Boat	108	52.0	8.5	54.7	9.4	49.3	9.6	51.8	8.9						
4. LSM Training	95	52.0	10.1	54.1	10.6	52.8	10.3	52.7	9.8						

[1] Scores expressed in Navy Standard Score form, M-50, $\sigma = 10$.

[2] Scores expressed in raw score form.

[3] Statistics for the parts of the Officer Classification Test were obtained from a sample of 74 cases. Statistics for the General Physics and General Mathematics Tests were obtained from a different sample of 120 cases.

[4] Scores for the parts of the Officer Classification Test were available for only 83 officers who had been assigned to NTSch (Tactical Radar) from Midshipmen and Indoctrination Schools. Scores on the CIC Aptitude Test were obtained on all men in 3 classes, N = 306.

ing was .54 for a sample of 120 officers. In this case the sum of scores on the General Physics and General Mathematics Tests correlated .60 with school success.

3. There is very little differentiation among these highly technical schools with respect to the sections of the Officer Classification Test which are most predictive. Superior performance on the Mechanical and Mathematical Sections appears to be necessary for success in practically all these schools. Special tests designed for prediction in each specific school might differentiate between them; the Officer Classification Test does not.

SEMI-TECHNICAL SCHOOLS. As indicated earlier, certain of the advanced schools require a lesser degree of skill and understanding in mathematics and the physical sciences than those listed as highly technical. The courses in the semi-technical schools involve two main elements: first, knowledge of the rudiments of navigation, with major emphasis on plotting; and second, rapid mastery of a large body of procedural doctrine, including terminology peculiar to the particular billet. This description certainly applies to the tactical radar, fighter director, and air combat intelligence billets, and at least the second aspect applies to communications training. (See Part II, of Table 1-XI.) Studies of these programs reveal a less definite pattern than studies of the more highly technical courses, but the following observations seem justified on the basis of the research in this area.

1. Success in the semi-technical advanced officer schools can be predicted reasonably well by using the Verbal and Mathematical Sections of the Officer Classification Test. For a sample of 83 officers in three classes at the Tactical Radar School, final grades correlated approximately .32 with the Verbal Section and .49 with the Mathematical Section of the Officer Classification Test. The study of officers in Air Combat Intelligence School provided evidence of a negatively accelerated nonlinear relationship between success in that school and scores on the Verbal Section of the test. It also yielded *r*'s of .42 and .52 between school grades and the Mathematical and Spatial Sections, respectively.

2. Special aptitude tests were developed which will predict school success reasonably well, but not appreciably better than selected parts of the Officer Classification Test. The CIC Aptitude Test yielded correlation coefficients of .54 with school success at NTSch (Tactical Radar), a figure comparable to that of the Mathematical Section of the Officer Classification Test. For a sample of 70 cases, application of the Wherry-Doolittle procedure yielded a multiple correlation coefficient of .58 when scores on all four sections of the Officer Classification Test and on the CIC Aptitude Test were combined. Although the sample was not adequate to make the foregoing analysis definitive, it tends to indicate that for a semi-technical school of this type, a standardized battery of apti-

tude tests for officers will probably yield results about as satisfactory as specialized aptitude tests designed to predict success in specific schools.

3. Previous training in science and mathematics seems to be desirable, although differences in favor of the officers so trained as compared to officers with non-technical training were not very large. Officers with scientific training (including the fields of engineering, physics, chemistry, and mathematics) made a mean final grade of 3.16 in the Tactical Radar School, while officers with non-scientific training (in any field other than those listed above) made a mean score of 3.08. The difference of .08 is approximately .4 of the standard deviation of either distribution. This difference was statistically significant but not large enough to be of much practical significance.

4. Age and civilian occupational experience appear to be of little predictive value in these schools. At the Tactical Radar School the younger officers made somewhat higher final grades than the older men, although between the ages of 22 and 30 the differences were very slight. Over 30, there appeared to be a definite reduction in mean final grade. No men under 22 were included in the study. Similar results were obtained in the Air Navigation School. In the Aviation Engineering School the older men had more training and experience than the younger men, hence a significant positive correlation was obtained (.24).

The following additional comments should be made concerning the researches conducted in semi-technical schools.

1. The Officer Personal Inventory and portions of the Biographical-Preference Inventory were tried out in two studies at semi-technical schools (Tactical Radar and Air Combat Intelligence) and were found to be of very little predictive value. The evidence is much too meager, however, to support a generalization regarding the potential usefulness of instruments of this type.

2. The aviation engineering training program provided an interesting verification of the general patterns described above. It appeared to call for a combination of the abilities necessary for the highly technical and semi-technical programs. The final grade was based on (1) performance in engineering classroom and shop work, and (2) performance in classes designed to train the officers for administrative work, with about equal nominal weight given to the engineering and administrative aspects of the course. In line with this analysis of course content, success in this school was best predicted by the Verbal and the Mechanical Sections of the Officer Classification Test (r's of .33 for each part). The group varied in amount of schooling from 10 to 18 years. Since years of schooling correlated .35 with average grades, a minimum requirement of at least one year of college seemed justified.

3. The correlations of the Mechanical and Mathematical Sections of the Officer Classification Test with success in Communications School are of some interest. These officers were assigned to this type of training in part because of their above average verbal ability. But among officers

of this level of verbal ability, those with high mechanical and mathematical scores surpassed those with low scores in these tests. Thus, it appears that fairly high scores on the Mechanical and Mathematical Sections of the test may be a reasonable requirement of a secondary nature for those officers who are assigned to Communications School.

The substance of the foregoing discussion is to place a good deal of stress on the importance of mechanical and mathematical ability in the officer population, not for just a few schools but for the majority of the training programs. Thus, it would seem justifiable in the future to establish procurement policies to attract a large proportion of officer candidates with mechanical and mathematical ability superior to the average of the officer population of World War II. There will apparently still be some value in distributing these officers among the various advanced schools in accordance with their particular patterns of abilities, but since so many schools apparently require fundamentally the same kinds of ability (particularly mathematical ability) and in approximately the same degree, it is highly important that the officer population include men well qualified in these respects.

Prediction of Success in Operational Training Programs

The data for the operational training programs are presented in Part III of Table 1-XI.

HIGHLY TECHNICAL TRAINING. The degree of technicality involved in the operational training program is directly a function of the complexity of the type of ship for which men are being prepared in the program. The submarine is a highly complex mechanism and the training program for submarines is correspondingly technical. Likewise, the training program for destroyer officers called for mastery of more technical material than did the corresponding training for amphibious craft and motor torpedo boats.

The following generalizations regarding the relatively technical operational training programs are based on studies of only these two programs. The sampling in each was fairly extensive, however, and in the case of the destroyer program at least, the criterion was perhaps the best available in any operational training program studied.

Success in the highly technical operational training program can be predicted fairly well on the basis of total Officer Classification Test scores with the Mathematics and Spatial parts of the test contributing most in predictive efficiency. Prediction is not as high, however, as in the advanced officer schools previously described. In a sample of 293 submarine officers, total scores on the Officer Classification Test correlated .40 with final class rank; the Mathematical

Section yielded a correlation of .34. Correlations with the pass-fail criterion were somewhat higher. At the destroyer school the Mathematical Section correlated .36 and the Spatial Section correlated .38 with the overall criterion.

Non-linear relationships between aptitude test scores and school success were found in the case of the destroyer officers, indicating negatively accelerated relationships, as in the case of the air combat intelligence group. The total Officer Classification Test scores yielded an *r* of .46 and an *eta* (final ratings on test scores) of .55.

Little relationship was found between age and school success. At the destroyer program, for example, a low non-linear relationship was observed between age and school success, with officers outside the preferred age range of 25 to 32 rated higher than those within the preferred age range.

Again in the case of the destroyers, mathematics-science-engineering training was found to be somewhat desirable, but civilian occupation apparently bore little relationship to success at the school. Officers with mathematics, science, and engineering training were found to be significantly superior to those whose college major came from the fields of literature, language, and social studies.

RELATIVELY NON-TECHNICAL TRAINING. This classification includes the programs for training LSM and motor torpedo boat officers. The outstanding characteristic of these programs which differentiates them from those previously described is the emphasis on practical afloat training and the handling of enlisted personnel. In each case, classes were held and grades assigned for classroom work, but there were practically no failures on academic grounds and the principal emphasis throughout was placed on the afloat training.

Civilian experience appears to be a more significant index of probability of success in the non-technical operational training programs than in the more technical schools. Managerial experience and experience with the handling of small boats were found to be definitely important factors in the selection of motor torpedo boat officers. Among LSM officers a difference was found in favor of those with mechanical experience, although the relationship was not statistically significant with the relatively small sample studied.

The Officer Classification Test was found to be of relatively little value in predicting success in these programs. Studies of motor torpedo boat officers showed almost no relationship between Officer Classification Test scores and final test grades, although moderately high relationships were found for specific courses in the introductory theory part of the training program (Table 1-XI). Among the LSM officers, the Verbal and Mechanical Sections of the Officer Classifi-

cation Test correlated .22 and .20 respectively with afloat ratings on the officer training courses (it should be noted also that these tests correlated .32 and .29 respectively with average grades in the classroom instruction phase of the LSM program).

Over a wide range from approximately 18 to 35 years, age was found to have very little relationship to success in these training programs. An exception to the above statement should be indicated in the case of a group of motor torpedo boat officers given ratings on the quality of their performance in combat; in this case the older men were rated somewhat superior to the younger, largely because of their ability to handle the administrative tasks involved in leadership of squadrons and larger units in motor torpedo boat warfare.

Summary

A dual summary will be presented. First, the principal findings of the research program will be summarized in terms of the type of training program, indicating the factor or factors which have been found to be most highly predictive of success in each type of training program. Secondly, consideration will be given to each type of predictive variable and an attempt made to indicate the kinds of advanced training programs, if any, for which such variable has been found to have predictive value. It should be recognized that generalization with respect to the efficacy of various predictive variables is necessarily limited by the fact that relatively few different kinds of predictors were used in the research program.

Advanced Schools. Aptitude test scores were found to be more predictive of success in advanced schools than age, education, civilian experience, and the other predictors studied. For all the advanced schools, the Mathematical Section of the Officer Classification Test was fairly predictive of success (r's with school grades in the neighborhood of .50). In schools whose curricula involved rapid learning of a large amount of procedural doctrine, the Verbal Section of the test also proved to be usefully predictive; but in the more strictly technical schools, the Mechanical Section was best used in conjunction with the Mathematical Section. With the possible exception of such highly technical schools as Pre-Radar (or schools such as the Sonar School, requiring exceptional sensory ability), it appears that a standardized battery of aptitude tests comparable to the Officer Classification Test will yield as satisfactory results in predicting school success as will more highly specialized aptitude tests designed specifically for predicting success in specific schools.

Operational Training Programs. In the more technical of the operational training programs, total score on the Officer Classification

Test proved to be more predictive than any other variables investigated (*r*'s of approximately .45 with final grades in the training programs), although college training in science and mathematics was also related to success. In the non-technical operational training programs, the test scores proved to be of very little predictive value, whereas civilian occupational experience was found to be of considerable significance. Particularly in a specialized but non-technical program, such as the training of motor torpedo boat officers, managerial experience in civilian life proved to be related to ratings on performance afloat, both under instruction and in combat. In general, it appears that as one goes from the more technical to the less technical programs, the predictive value of aptitude test scores decreases and the significance of civilian training and occupational experience increases.

SUMMARY OF FINDINGS BY PREDICTIVE VARIABLE. In interpreting the following summary, it should be remembered that the conclusions are based on conditions operative during World War II. If the basic officer population were selected in a significantly different manner or if the curricula and grading policies of the schools were changed considerably, these generalizations might not be applicable.

Age. The age variable was investigated in almost all of the training programs and proved to be of little importance for predicting success in most of the officer programs. It may well be that maturity is a significant factor, as is claimed by a large portion of school officers. If so, age is apparently an unreliable measure of maturity.

Education. Amount of civilian education, as measured by the number of years spent in school, is not ordinarily important among a group of officers all of whom have had some college training. But it may be significant if a wider range of education is represented, as in some of the special aviation training programs where some of the enrollees had less than college training, while others had done graduate work. Type of education, as indicated by major field of of study in college, was found to have some predictive value for advanced schools and for the more technical of the operational training programs.

Civilian Occupational Experience. Although in almost all schools officers with experience in mechanical and engineering occupations proved to be slightly more successful than officers who had done non-technical work in clerical and sales occupations, this factor was generally of less predictive value than the aptitude tests. In the relatively non-technical operational training programs, however, where handling of enlisted men and acceptance of administrative responsibility overshadow mastery of technical subject-matter, civilian occupational experience of an engineering or managerial type

apparently bears more relationship to success in the training program than does proficiency on standardized aptitude tests.

Aptitude Test Scores. 1. General aptitude tests. The Officer Classification Test proved to be reasonably predictive in all except the non-technical operational training programs. The type of program in which each section of the test was most predictive is indicated below:

Verbal—to a moderate degree in all the advanced training programs; highest average scores were made by officers in the semi-technical advanced schools.

Mechanical—the highly technical schools such as pre-radar, ordnance, and gunnery, advanced fire control, and diesel engineering.

Mathematical—all the advanced schools, and the operational training programs, with the exception of LSM afloat training. Mathematical ability is apparently the basic ability most in demand in advanced training of naval officers.

Spatial—Air combat intelligence in the semi-technical advanced school group and the highly technical operational training programs (notably the destroyer and submarine programs).

2. Special aptitude tests. Tests of special aptitude proved to be of some value for prediction in the advanced officer schools, but they were only slightly more effective than the various parts of the Officer Classification Test.

Other Data. The scores on the Officer Personal Inventory were found to bear little relationship to school success; but it should be noted that in the instances where the Personal Inventory was studied (submarines, destroyers, tactical radar), the officers were so highly selected that there were very few cases with high Personal Inventory scores. With respect to the Biographical-Preference Inventory, duty preference statement, grades in basic training, interest inventories, self-evaluation, and evaluation by peers, so little evidence was obtained (in some cases none at all) that no generalizations are warranted.

In summarizing the findings on the prediction of success in advanced officer training programs it seems safe to conclude that tests such as the Officer Classification Test are useful in the classification of officer personnel. Other predictive indices are useful in specific situations, but none is universally applicable. Perhaps the most serious gap in the research completed to date is that pertaining to the measurement of personality and interest factors, and use of personality and interest measures in the prediction of success in training and on the job. By studying these factors and by making systematic coordinated studies in place of the piecemeal studies carried out during the war, the efficiency of prediction in advanced officer training programs should be considerably improved.

CHAPTER XII

PREDICTION OF SUCCESS IN ELEMENTARY SCHOOLS FOR ENLISTED PERSONNEL

THE selection of qualified trainees for enlisted elementary schools constituted one of the Navy's major personnel problems during World War II. As new schools were established, selection requirements had to be formulated for the guidance of the classification officers who were responsible for making school assignments; and revisions in these requirements had to be made as dictated by experience and Navy needs. The studies reported in this chapter are concerned with (1) determining the relationship between certain data on the Enlisted Personnel Qualifications Card and grades received in schools, and (2) establishment of the most suitable cutting scores on tests and minimum standards of personal requirements for different types of elementary schools.

Before the introduction of the Basic Test Battery in June 1943, the problem of the utilization of test scores in formulating selection requirements for elementary enlisted schools had been approached by studying the relationship between scores on the old Navy test battery and school grades. Selection requirements involving the new Basic Test Battery were tentatively established on the basis of correlation coefficients computed between the tests previously in use and those newly adopted. In the determination of selection criteria for some of the schools, the new tests were administered experimentally to classes just beginning their instruction in the school and also to classes ready for graduation. Tests which showed the best predictive powers were selected for incorporation into the selection requirements, and cutting scores on the tests were set. Since recruits with low scores on the selected tests appeared to have little chance of completing successfully a school course, they were not to be recommended. The main factors basic to the determination of the level of the cutting scores were the difficulty of the school curriculum and the quota of personnel to be assigned to the school.

From the data gathered upon recruits assigned before the new tests were in use and from data on those assigned on the basis of tentative selection requirements involving the new test battery, it was possible tentatively to determine the effectiveness of the new selection criteria. A more comprehensive evaluation, however, awaited the gathering of data on large numbers of recruits who were assigned to schools with Basic Test Battery scores as part of the information available at the time of assignment.

The Populations Studied

The school groups used in the analyses to be presented were tested with either Form 1 of the basic battery tests in 1943 or with Form 2 in 1944. Data on trends, which have been presented in Chapter VI, indicate that the time of the year in which school assignments are made constituted one of the significant factors in the determination of the type of personnel assigned to school; a second important factor was the changes in school quotas. The subjects taking Form 1 of the battery were tested between June 1943 and January 1944 with approximately the same number of trainees per month. A large proportion of the subjects who were given Form 2 of the battery were tested in the months of June and July when the mean monthly scores on the tests were among the highest for the year.

In the assignment of personnel by classification interviewers at the training centers, test scores have constituted only a part of the information which is used in determining whether a recruit is to be sent to a school or assigned to "general detail". Other data include: (a) educational background, (b) previous experience, (c) interests, and (d) an overall appraisal by the classification interviewer. The complete classification procedure is described in Chapter III.

In selecting schools for validity studies, first consideration was given to those schools in which attrition was high and for which the quotas were large. Data were obtained for a number of classes for each type of school studied; several school locations also were represented for each type. One factor in the selection of classes was that of securing a sufficiently large population in the class to lend stability to the statistics obtained.

Validity data will be presented for the schools listed below. The abbreviations in most instances are those of the Navy ratings for which the schools trained. Basic engineering schools did not train for a single specific rating, hence the abbreviation B. Eng. Hospital corps schools trained both hospital apprentices and pharmacist's mates and are here designated HC.

Aviation Machinist's Mates	AMM
Aviation Ordnancemen	AOM
Aviation Radiomen	ARM
Basic Engineering	B. Eng.
Diesel	MoMM
Electrical	EM
Fire Controlmen	FC
Fire Controlmen (O)	FC (O)
Gunner's Mates	GM

Hospital Corps	HC
Hospital Corps (WR)	HC (WR)
Machinist's Mates	MM
Quartermasters	QM
Radio	RM
Radar Operators	RdM
Signal	SM
Storekeepers	SK
Storekeepers (WR)	SK (WR)
Torpedomen	TM
Yeomen	Y
Yeomen (WR)	Y (WR)

Predictive Factors

The principal selection factors studied include scores on six general tests of the Basic Test Battery and three special tests of the battery, a quality classification assigned to each man by the classification interviewer, age, and amount of civilian education. The general tests, together with the appropriate abbreviations are:

General Classification Test	GCT
Reading Test	READ
Arithmetical Reasoning Test	ARI
Mechanical Aptitude Test	MAT
Mechanical Knowledge Test (Mechanical Score)	MKM
Mechanical Knowledge Test (Electrical Score)	MKE

The special tests are:

Spelling Test	SPELL
Clerical Aptitude Test	CLER
Radio Code Test—Speed of Response	CODE

A description of the first eight of these tests has been given in Chapter VI. The Radio Code Test has been described in Chapter VIII.

While statistical data on Forms 1 and 2 of the six general tests will be presented, data for the other three are available for Form 1 alone. The classes of the schools for which data on the special tests are presented are identical with those used in the analyses of Form 2 of the six general tests. This is because the special tests did not become a part of the testing program at the naval training centers until January 1944, at which time Form 2 of the general tests was introduced.

The "quality classification" constitutes a recommendation by the interviewer relative to the recruit's qualification for the school for which he is recommended. Quality "1" indicates exceptional background qualifications for a particular school assignment although

not necessarily characterized by high test scores; quality "2" indicates high scores and no unique occupational background; and "3" indicates that the man has only limited school possibilities. Men not assigned a quality classification were recommended for "general detail."

The statements made by the recruit relative to his age and highest school grade completed as recorded on the Qualifications Card were used.

Criterion

The criterion of achievement in all cases was the final school grade assigned to the trainee, by whatever process the school arrived at it. At the time Form 1 was administered, these grades were determined from marks on written tests constructed in the various schools and marks assigned in laboratory or practical courses. Standardized achievement examinations, both written and performance, were in use in some of the schools for which Form 2 data are reported. When they were used, they were to be given a crude weight of approximately one-third in making up the final grade. Standardized written achievement examinations and performance tests were used in the gunner's mate, radar operator (some schools), signal, torpedoman, and yeoman schools. Standardized written achievement examinations but not standardized performance tests were used in fire controlman, storekeeper, and electrical schools. No standardized written or performance tests were used in the aviation machinist's mate, aviation ordnanceman, aviation radioman, fire controlman (O), hospital corps, quartermaster, and radio schools.

The exact nature of the criterion varied from school type to school type, from location to location, and from class to class. Efforts were exerted to raise the reliability of the school grades and to increase their relationship to the actual jobs the graduates were expected to perform, so that the criterion data for the later classes, the classes which had been tested on Form 2 of the Basic Test Battery, are probably more reliable and more valid.

Statistical Procedures

The basic technique of statistical analysis of the effectiveness of tests was the Pearson product-moment coefficient of correlation, computed from grouped data. The predictive value of a test was judged in terms of the coefficient of correlation between the test scores and final grades; this coefficient is usually referred to hereafter as the "validity coefficient."

As stated earlier, the tests of the Basic Test Battery were used for

selection purposes before validation studies could be made. For each type of school the test or tests were specified and cutting scores set on the basis of previous general experience and rational considerations. For this reason, the graduates of a type of school were more homogeneous with regard to test scores than was the general recruit population from which selection was made. The effect of this restriction in range of test scores was to lower the validity coefficients. In general, the decreased variability was most pronounced in the most "plausible" test for a given school, so that the test which was presumably the best might be made to appear among the worst. To facilitate comparison of validity coefficients, Kelley's formula [1] for correction for restriction in range was used. This formula makes possible an estimate of what the validity coefficient would have been if the general recruit population had been sent to the school rather than the actual restricted sample. The estimated or corrected coefficients (r_c) were, in practice, obtained from a table constructed to represent Kelley's formula (Appendix E-4).

The formula assumes (1) that the restriction in the criterion is produced entirely by selection on the given test, (2) that the regression of the criterion on the test is rectilinear, and (3) that the chart representing the correlation between the test and the criterion is homoscedastic. In the present case, as indicated below, these three assumptions are not perfectly fulfilled.

The selection was made by classification interviewers using all information available, including test scores and other factors, so that presumably the criterion was restricted also by selection on factors in addition to the test under consideration. Only the grades of graduates have been studied; if those for dropped trainees had been included, the criterion would have been less homogeneous. The effect of this additional restriction on the criterion is, in general, to lower the apparent validity of the test. Furthermore the other selection factors (in addition to test scores) tend to be used in such a way that deficiencies in one qualification are compensated for by higher standing in another. For example, if a cutting score of 50 were established as a qualification for a given school, a man with a score of 47 would be accepted only if he were unusually well qualified by previous background or if his other test scores were

[1] $$r_c = \frac{r}{\sqrt{r^2 + \frac{\sigma^2}{\Sigma_2} - \frac{r^2\,\sigma^2}{\Sigma_2}}}$$

where r_c = the correlation coefficient corrected for restriction in range,
r = the obtained correlation coefficient,
σ = the standard deviation of the test in the restricted sample,
Σ = the standard deviation of the test in the unrestricted population.

considerably higher. If, in general, two selection devices are so used that men who are low in one are accepted only if they are high in the other, the correlation coefficient between the two devices in the selected group is systematically lowered, and the apparent validity of each device is also lowered more than the mere restriction in range would indicate. The general effect, then, of using multivariate selection rather than the single variate selection assumed by the formula, is to produce an underestimate of validity, and the values reported in this study are probably systematically too low.

Violations of the assumptions of rectilinear regression and homoscedasticity can produce either underestimates or overestimates. If the line representing regression of the criterion on the test is negatively accelerated, or if men with high test scores are more variable on the criterion than those with low test scores, the result will be again an underestimate of what the correlation between the test and the criterion would have been if an unselected recruit population had been sent to school. A positively accelerated regression line or larger criterion variation for low test scores would result in an overestimate. The negatively accelerated curve, following the general principle of diminishing returns, is more likely in a situation of this sort.

Another effect of selection upon the obtained validity coefficients is not corrected by Kelley's formula. If the regression is curvilinear, a selected group with a mean test score higher than that of the recruit population but with as large a standard deviation will show a lower validity if the regression line is negatively accelerated and a higher one if it is positively accelerated. In general, insufficient data are available to determine the shape of the regression line in an unselected group, so that no correction can be made except to note that the reported validity coefficients are probably underestimates of the true values.

In addition to the technique of correlation, the comparison of means and standard deviations for the graduates of various schools sheds some light on the relation between test scores and success in schools. These means, of course, reflect primarily the selection procedures in operation at the time; but to the extent that trainees lacking the requisite ability were not graduated, the data give some evidence of validity and help to determine appropriate standards of test performance.

For determining the usefulness of combinations of tests for predicting school success, the conventional procedure is determination of multiple correlation coefficients. This technique arrives at the combination of tests which will give the best possible prediction and gives the exact weight to be assigned to each test. For adminis-

trative reasons, however, fractional weights for tests were not considered practical in actual field operations; therefore the test combination was selected for which the simple sum of the scores had the best predictive efficiency. With a test battery in which the test intercorrelations are relatively high and all tests are expressed in standard scores, the two techniques are almost equivalent; they usually result in the same combinations of tests and indicate approximately the same predictive efficiency. The reason for this is that with high inter-test correlation coefficients, if the second test adds much to the usefulness of the first test, the weights obtained by the multiple correlation technique are usually almost equal. The formula for obtaining the correlation coefficient of the summed score of two tests with the criterion, when both tests have equal standard deviations, is:

$$r_{(a+b)x} = \frac{r_{ax} + r_{bx}}{\sqrt{2\,(1 + r_{ab})}}$$

where a is the first test,
b is the second test,
and x is the criterion;
r_{ax} is the validity of the first test,
r_{bx} is the validity of the second test,
and r_{ab} is the correlation between the tests.

In this study the test intercorrelations used were those obtained on several samples of the recruit population (Chapter VI). The test intercorrelations for each class were not computed. The effect of selecting trainees on the basis of a combination of two or more tests is to lower the test intercorrelations in the selected group more than the restriction in range of either test would indicate. If, for example, a man with a low score on the General Classification Test is assigned to school only provided he has a high score on the Arithmetical Reasoning Test, and vice versa, the result is a disproportionate number of people in the selected group who are high on one test and low on the other. The effect of this phenomenon is to lower the test intercorrelations in the selected group. In the few school classes where the test intercorrelations were computed, they were found to be markedly low.

Since the standard deviation of the sum of two tests is given by the formula, $\sigma_{(x+y)} = \sqrt{\sigma_x^2 + \sigma_y^2 + 2r_{xy}\sigma_x\sigma_y}$, the use of a value for r_{xy} that is too high will give a spuriously high value for the variability of the sum of two tests in a school class. When the formula for correction for restriction in range is used, the correction obtained is too low; and the net result is an underestimate of what

the correlation coefficient between the criterion and the sum of two tests would have been if the unselected recruit population had been sent to school.

Findings

BASIC TEST BATTERY VALIDITY DATA. In the presentation of the validity data, the elementary schools studied have been divided into three groups. This grouping is partly on the basis of the results of the validation study, but it follows closely a rational classification made before the Basic Test Battery was constructed.

Group 1 Schools. The schools included in Group 1 train personnel for the following ratings: aviation ordnanceman, aviation radioman, fire controlman, fire controlman (O), hospital corps, quartermaster, radarman, radioman, signalman, storekeeper and yeoman. The General Classification Test and the Arithmetical Reasoning Test tend to be most useful in predicting success in Group 1 schools, which are characterized by stress on verbal material and the ability to use symbols.

The validity coefficients for Group 1 schools are shown in Table 1-XII. The figures are the simple weighted means of the validity coefficients for the classes of each school type. Separate data have been presented for Form 1 and Form 2, not primarily because the two forms of the tests are so much different, but because during the time interval involved, a number of changes took place in the curricula of the schools, in instructional methods, in grading practices, and in the standards used in selecting trainees.

The General Classification Test, Reading Test, and Arithmetical Reasoning Test all show high predictive efficiency for this group of schools. Taking the data for both forms together, the median corrected validity coefficient for the General Classification Test is .50; for the Reading Test .46; for the Arithmetical Reasoning Test .47. The three mechanical tests have much lower validity coefficients for these schools, with the median corrected validity coefficient at .31 for the Mechanical Aptitude Test, .21 for the Mechanical Knowledge Test (Mechanical Score), and .30 for the Mechanical Knowledge Test (Electrical Score).

In comparison with Form 1, the corrected correlation coefficients for Form 2 show, in general, an increase in the validity of the General Classification Test averaging about .055, a decrease for the Reading Test of .02 and a decrease of about .07 for the Arithmetical Reasoning Test.

The test means and standard deviations (expressed in Navy Standard Scores) for Group 1 schools are shown in Table 2-XII. There is evidence that the graduates have been selected on the basis of the

TABLE 1-XII. Group 1 Schools: Correlation coefficients (raw and corrected for restriction in range)[1] between scores on six tests of the Basic Test Battery and final school grades for graduates of fourteen types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test Form	Test: GCT r	GCT r_c	READ r	READ r_c	ARI r	ARI r_c	MAT r	MAT r_c	MKM r	MKM r_c	MKE r	MKE r_c
Aviation Ordnancemen	4	469	1	.34	.43	.36	.47	.40	.52	.42	.52	.34	.36	.35	.40
	5	805	2	.35	.53	.31	.39	.35	.37	.22	.28	.28	.37	.38	.48
Aviation Radiomen	2	247	1	.33	.49	.33	.43	.27	.34	.21	.27	.25	.29	.29	.35
	5	737	2	.34	.54	.36	.47	.39	.45	.27	.37	.18	.22	.30	.29
Fire Controlmen	5	449	1	.43	.52	.41	.50	.43	.55	.31	.41	.16	.17	.17	.19
	6	478	2	.34	.57	.35	.47	.38	.49	.24	.34	.26	.33	.38	.54
Fire Controlmen (O)	5	370	2	.29	.51	.32	.43	.34	.40	.28	.41	.18	.21	.22	.30
Hospital Corps	4	1,188	1	.33	.47	.29	.40	.28	.37	.23	.29	.16	.19	.24	.30
	6	409	2	.61	.72	.54	.59	.59	.56	.42	.46	.23	.25	.43	.50
Hospital Corps (WR)	4	594	2	.48	.62	.46	.52	.44	.46	.20	.29				
Quartermasters	13	1,321	1	.34	.42	.34	.45	.38	.52	.23	.30	.11	.11	.10	.12
	12	470	2	.27	.48	.31	.46	.42	.41	.24	.30	.09	.08	.22	.18
Radar Operators	5	1,158	1	.48	.60	.53	.67	.51	.61	.42	.49	.32	.34	.33	.37
	5	402	2	.44	.65	.40	.52	.43	.48	.37	.50	.30	.36	.42	.50
Radio	10	1,542	1	.26[2]	.35[2]	.24	.32	.24	.32	.15	.18	.05	.07	.08	.10
	16	1,013	2	.15	.23	.15	.23	.22	.23	.09	.12	.13	.13	.16	.18
Signal	18	2,496	1	.32	.44	.30	.41	.35	.43	.19	.23	.06	.07	.08	.10
	11	864	2	.24	.44	.25	.32	.35	.37	.26	.32	.11	.13	.25	.32
Storekeepers	5	661	1	.41	.49	.32	.40	.42	.53	.25	.28	.20	.21	.23	.25
	4	368	2	.40	.58	.38	.47	.47	.54	.28	.33	.12	.14	.25	.32
Storekeepers (WR)	4	678	1	.34	.42	.34	.46	.50	.59	.28	.36				
	2	241	2	.33	.52	.25	.36	.33	.44	.10	.17				
Yeomen	2	295	1	.42	.49	.37	.48	.49	.60	.21	.25	.28	.31	.20	.24
	7	325	2	.30	.47	.36	.47	.33	.35	.24	.32	.10	.11	.13	.16
Yeomen (WR)	2	738	1	.52	.63	.46	.59	.53	.62	.30	.38				
	6	1,190	2	.44	.65	.48	.58	.47	.50	.29	.38				

[1] Kelley, Truman L. *Statistical Method.* New York: Macmillan, 1924, p. 225.
[2] Number of classes = 26, number of trainees = 3,339.

TABLE 2-XII. Group 1 Schools: Means and standard deviations on six tests of the Basic Test Battery for graduates of fourteen types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test Form	Test											
				GCT		READ		ARI		MAT		MKM		MKE	
				M	σ	M	σ	M	σ	M	σ	M	σ	M	σ
Aviation Ordnancemen	4	469	1	52.0	6.5	52.9	7.1	51.0	7.1	54.4	7.6	52.3	9.5	53.1	8.5
	5	805	2	60.3	5.9	62.9	7.8	61.7	9.2	60.9	7.3	57.8	7.3	58.3	7.3
Aviation Radiomen	2	247	1	57.2	6.1	53.3	7.0	50.7	7.7	54.6	7.3	50.3	8.6	51.3	8.3
	5	737	2	62.8	5.6	65.7	7.2	64.8	8.4	62.4	6.9	52.8	7.9	58.0	7.4
Fire Controlmen	5	449	1	58.9	7.8	54.4	7.7	53.8	7.1	54.8	7.4	55.5	9.4	58.9	9.0
	6	478	2	64.3	5.0	66.8	7.0	66.5	7.3	62.6	6.8	61.5	7.4	65.5	6.3
Fire Controlmen (O)	5	370	2	62.4	4.8	65.2	6.9	62.1	8.2	59.8	6.5	54.2	8.7	57.7	7.0
Hospital Corps	4	1,188	1	54.8	6.5	50.4	6.9	48.3	7.4	50.0	7.8	47.0	8.6	49.4	7.8
	6	409	2	61.5	7.5	62.2	8.8	60.3	10.8	57.6	8.7	50.5	9.1	56.8	8.2
Hospital Corps (WR)	4	594	2	57.1	6.9	56.4	8.7	49.0	9.7	48.1	7.1				
Quartermasters	13	1,321	1	62.1	7.6	57.5	7.0	56.0	7.0	56.0	7.7	52.5	9.0	53.4	8.8
	12	470	2	65.8	4.6	67.8	6.2	67.5	7.0	61.6	7.3	54.3	9.0	59.3	7.2
Radar Operators	5	1,158	1	57.8	7.4	51.5	7.0	51.6	7.5	51.8	8.1	53.4	9.4	54.1	8.9
	5	402	2	59.5	5.7	61.5	7.2	59.6	8.4	57.9	6.7	53.9	8.1	56.2	8.0
Radio	10	1,542	1	52.8[1]	7.2[1]	50.0	7.1	48.5	7.6	49.5	8.1	45.8	8.8	49.4	8.9
	16	1,013	2	61.4	5.9	63.0	7.8	60.8	9.6	58.5	7.4	51.8	8.6	56.9	8.3
Signal	18	2,496	1	52.6	6.8	49.9	7.2	47.9	7.9	49.8	8.1	45.6	8.5	47.5	7.9
	11	864	2	59.3	5.4	63.6	7.6	61.4	9.4	58.2	7.7	49.4	8.3	54.8	7.5
Storekeepers	5	661	1	58.2	8.1	52.1	7.7	53.4	7.5	51.0	8.6	50.7	9.4	51.9	9.4
	4	368	2	63.5	6.0	63.9	7.6	65.1	8.1	57.9	8.1	53.8	9.3	56.7	7.6
Storekeepers (WR)	4	678	1	59.3	7.5	52.2	6.8	48.7	7.8	43.9	7.4				
	2	241	2	62.6	5.7	62.8	6.7	61.2	7.1	52.1	6.2				
Yeomen	2	295	1	57.6	8.4	52.5	7.4	50.9	7.4	50.0	8.4	44.5	8.9	46.9	8.1
	7	325	2	61.8	5.5	63.0	7.3	61.4	9.2	56.8	7.5	48.0	9.2	53.5	8.1
Yeomen (WR)	2	738	1	55.0	7.7	48.8	7.2	43.1	7.8	40.4	7.6				
	6	1,190	2	58.4	5.6	57.5	7.6	50.4	9.3	48.2	7.3				

[1] Number of classes = 26, number of trainees = 3,339.

tests or on variables correlated with the tests, since most of the means are above 50 and most of the standard deviations are below 10. The rigid use of a cutting score of 50 would yield a mean of 58 and a standard deviation of 6 in the selected group.

This degree of selection is approached on Form 1 of the General Classification Test in all schools except those for aviation ordnancemen, radiomen, and signalmen. Except for quartermaster schools, Form 1 of the Reading Test and Arithmetical Reasoning Test shows much less evidence of selection.

For two of the fourteen schools in Group 1, Fire Controlman (O) and Hospital Corps (WR), data were available for only one of the forms. In the twelve schools where data are available for both forms, the increase in the mean score from Form 1 to Form 2 is 5.2 for the General Classification Test, 11.3 for Reading Test, and 11.4 for Arithmetical Reasoning Test. The standard deviations for Form 2 are 1.6 lower for the General Classification Test, 0.2 higher for Reading Test, and 1.2 higher for Arithmetical Reasoning Test.

Three factors enter into the differences between the mean scores on the two forms. (1) As the classification program became more definitely crystallized, the degree of selection in terms of test scores probably increased. The lower standard deviations for Form 2 of the General Classification Test point in that direction. (2) The time of year in which the Form 2 data were collected for this study would also point toward higher means, since the trainees involved in the study reported here were inducted during the May-August period for which mean test scores for the entire recruit population run high (Chapter VI). (3) Finally, there are some differences in the calibration of scores on the tests of Form 2, so that a given group will obtain somewhat higher means and standard deviations on Form 2 than on Form 1, especially for the Reading Test and the Arithmetical Reasoning Test.

Although the mean scores are higher on Form 2 for all the schools where comparison is possible, the relative standards for the different schools remain about the same. The rank-difference correlation coefficients between Form 1 and Form 2 means for the twelve schools are about .80 for the General Classification Test, Arithmetical Reasoning Test, or the sum of the two.

In general the averages of the mean scores on the six tests are high for quartermaster, fire controlman, storekeeper, and aviation radioman schools. For radio, hospital corps, signal, and yeoman schools, these averages are low. The other schools fall in between.

The validity coefficients of the summed scores of indicated tests for Group 1 schools are shown in Table 3-XII. The addition of a second test to the "best" single test sometimes raises and sometimes

lowers the validity coefficient, with the overall result being a small increase. The multiple correlation coefficient for the same combination of tests scores is only slightly higher than the validity coefficient of the unweighted sum. The slight gain in validity coefficient probably would not, in general, justify the addition of a second test to the first. But less shrinkage of validity in future samples may be expected from using the summed scores of two tests than from using

TABLE 3-XII. Group 1 Schools: Correlation coefficients between sums of scores on two tests of the Basic Test Battery and final school grades for graduates of fourteen tyes of elementary Naval Training Schools

Naval Training School	Test Form	Pair of Tests Whose Summed Score Has Highest Validity	Validity Coefficient Based on Summed Score	Multiple *R*	Validity Coefficient for GCT and ARI
Aviation Ordnancemen	1	ARI + MAT	.58	.58	.54
	2	GCT + MKE	.55	.55	.48
Aviation Radiomen	1	GCT + MKE or READ	.48	.50	.45
	2	GCT + READ	.53	.54	.52
Fire Controlmen	1	GCT + ARI	.58	.58	.58
	2	GCT + MKE	.60	.60	.56
Fire Controlmen (O)	2	GCT + READ	.49	.51	.48
Hospital Corps	1	GCT + READ or ARI	.46	.47	.46
	2	GCT + READ or ARI	.68	.72	.68
Hospital Corps (WR)	2	GCT + READ	.59	.62	.57
Quartermasters	1	READ or GCT + ARI	.51	.53	.51
	2	GCT + READ	.49	.49	.47
Radar Operators	1	READ + ARI	.67	.70	.66
	2	GCT + MAT	.63	.66	.60
Radio	1	GCT + ARI	.36	.36	.36
	2	READ + ARI	.25	.25	.24
Signal	1	GCT + ARI	.47	.47	.47
	2	READ + ARI	.43	.44	.43
Storekeepers	1	GCT + ARI	.55	.56	.55
	2	GCT + ARI	.59	.60	.59
Storekeepers (WR)	1	READ + ARI	.57	.59	.55
	2	GCT + ARI	.51	.52	.51
Yeomen	1	GCT + ARI	.59	.61	.59
	2	GCT + READ	.49	.49	.43
Yeomen (WR)	1	GCT + ARI	.68	.68	.68
	2	GCT + READ	.64	.65	.61

scores on the "best" single test, since a broader predictive base is used.

The schools in Group 1 are characterized by the frequency with which the sum of the General Classification Test plus the Arithmetical Reasoning Test occurs as the best predictive combination. If, for purposes of administrative simplicity, it is desired to assign men to this group of schools as a group, the use throughout of the

General Classification Test plus the Arithmetical Reasoning Test would result in an average loss of validity of less than .02. Success in Group 1 schools can be predicted sufficiently well for classification purposes using the sum of scores on the General Classification Test and Arithmetical Reasoning Test as the predictive index. The chief exceptions are the radio schools, for which the Radio Code Test has been developed.

Group 2 Schools. Group 2 schools have in common an emphasis on mechanical subjects, together with some technical informational content. Included are the basic engineering schools and the schools for training aviation machinist's mates. The validity coefficients are shown in Table 4-XII. All of the basic tests of the battery show substantial validity coefficients but the three mechanical tests and the Arithmetical Reasoning Test run higher than the General Classification Test and Reading Test.

In comparison with Form 1, Form 2 validity coefficients are higher in the mechanical tests and somewhat lower in the Reading Test and Arithmetical Reasoning Test. In the case of the basic engineering schools, this change can be definitely attributed to a modification in the grading system whereby the shop subjects were given greater relative emphasis in determining school grades at the time the Form 2 data were obtained; a similar change in the procedures used in grading may also have taken place in the electrical schools.

The means and standard deviations for Group 2 schools are shown in Table 5-XII. The small number of school types does not warrant detailed comparison of test means, but aviation machinist's mate, electrical, and diesel graduates average somewhat higher than basic engineering and machinist's mate graduates on the Arithmetical Reasoning Test and the Mechanical Knowledge Test.

Table 6-XII shows validity coefficients for selected combinations of tests for Group 2 schools. The best single combination for the group as a whole is that of the Arithmetical Reasoning Test with either the Mechanical or Electrical Score of the Mechanical Knowledge Test—the Electrical Score for the electrical schools, the Mechanical Score for the others. The validity coefficients for the combinations of test scores range from .48 to .65, indicating satisfactory predictive efficiency. In Forms 4 and 5 of the Basic Test Battery, good results may be expected for Group 2 schools from the sum of the proposed Arithmetic Test and Mechanical Test, since the former test is to include arithmetical reasoning and arithmetical computation items, and the latter test is to include mechanical comprehension, and both mechanical and electrical knowledge items.

Group 3 Schools. While Group 1 schools involve a minimum of mechanical content and Group 2 schools combine mechanical and

TABLE 4-XII. Group 2 Schools: Correlation coefficients (raw and corrected for restriction in range)[1] between scores on six tests of the Basic Test Battery and final school grades for graduates of five types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test Form	Test: GCT r	GCT r_c	READ r	READ r_c	ARI r	ARI r_c	MAT r	MAT r_c	MKM r	MKM r_c	MKE r	MKE r_c
Aviation Machinist's Mates	5	750	2	.26	.44	.25	.31	.27	.31	.36	.48	.46	.64	.41	.56
Basic Engineering	13	2,255	1	.38	.51	.38	.48	.50	.60	.43	.53	.43	.47	.37	.42
	5	1,177	2	.35	.50	.36	.42	.45	.50	.38	.52	.45	.62	.38	.50
Diesel	9	2,160	1	.31	.42	.27	.35	.28	.38	.20	.25	.36	.42	.40	.46
Electrical	25	3,418	1	.40	.52	.39	.49	.46	.57	.33	.41	.34	.39	.42	.49
	18	1,331	2	.33	.50	.31	.38	.46	.51	.38	.50	.27	.36	.39	.57
Machinist's Mates	7	934	1	.27	.37	.22	.29	.39	.49	.37	.46	.39	.46	.36	.44

[1] Kelley, Truman L. *Statistical Method.* New York: Macmillan, 1924, p. 225.

TABLE 5-XII. Group 2 Schools: Means and standard deviations on six tests of the Basic Test Battery for graduates of five types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test Form	Test: GCT M	GCT σ	READ M	READ σ	ARI M	ARI σ	MAT M	MAT σ	MKM M	MKM σ	MKE M	MKE σ
Aviation Machinist's Mates	5	750	2	62.3	5.6	66.1	7.8	64.5	8.7	63.6	6.9	63.5	6.2	62.2	6.6
Basic Engineering	13	2,255	1	51.2	7.0	47.0	7.4	47.0	7.6	51.5	7.8	54.4	9.8	53.6	8.3
	5	1,177	2	58.6	6.0	60.4	8.1	58.2	8.6	59.1	6.6	60.8	6.4	57.5	7.0
Diesel	9	2,160	1	54.3	7.2	50.0	7.8	49.9	7.5	54.1	8.0	57.0	8.6	56.3	8.2
Electrical	25	3,418	1	53.3	7.1	50.3	7.4	49.1	7.6	52.9	7.8	54.7	8.5	58.8	8.2
	18	1,331	2	60.2	5.8	62.6	7.7	60.9	8.6	59.7	7.1	58.3	7.2	64.1	5.9
Machinist's Mates	7	934	1	52.0	6.8	48.1	7.2	49.2	7.4	52.5	7.5	55.8	8.1	54.7	7.9

symbolic or theoretical material, Group 3 schools approach purely mechanical assembly-disassembly content. Included in this group are the gunner's mate and torpedoman schools. The test validity coefficients for these schools are given in Table 7-XII. On Form 1 both parts of the Mechanical Knowledge Test predict fairly well, but the General Classification Test and Reading Test also make a substantial contribution. On Form 2, used at a time when the grading system probably placed somewhat greater emphasis on the shop subjects, the Mechanical Knowledge Test scores show higher validity, while validity coefficients for the General Classification Test and Reading Test decrease in magnitude.

The test means and standard deviations in Table 8-XII show the torpedoman school graduates to score higher than the graduates of gunner's mate schools, especially on both parts of the Mechanical

TABLE 6-XII. Group 2 Schools: Correlation coefficients between sums of scores on two tests of the Basic Test Battery and final school grades for graduates of five types of elementary Naval Training Schools

Naval Training School	Test Form	Pair of Tests Whose Summed Score Has Highest Validity	Validity Coefficient Based on Summed Score	Multiple *R*	Validity Coefficients for ARI + MKM or MKE
Aviation Machinist's Mates	2	MKM + MKE	.65	.65	.55
Basic Engineering	1	ARI + MKM	.64	.65	.64
	2	ARI + MKM	.65	.66	.65
Diesel	1	GCT + MKE	.50	.50	.48
Electrical	1	ARI + MKE	.62	.62	.62
	2	ARI + MKE	.60	.60	.60
Machinist's Mates	1	MAT + MKM	.52	.52	.49

Knowledge Test. In comparison with the other schools, evaluating each school in terms of the tests most appropriate for it, the torpedoman schools show test standards somewhat higher than average, while the gunner's mate schools are considerably below average.

Table 9-XII shows the validity coefficients of selected test combinations for Group 3 schools. Both schools show the same pattern: for Form 1, the best combination was Reading Test plus the Mechanical Score on the Mechanical Knowledge Test, while for Form 2 the best combination was the Mechanical Score plus the Electrical Score on the Mechanical Knowledge Test. The sum of the two parts of the Mechanical Knowledge Test is a satisfactory combination and it is probable that the Mechanical Test of Forms 4 and 5, which consists of mechanical comprehension, mechanical knowledge, and electrical knowledge items, will be adequate for the selection of

TABLE 7-XII. Group 3 Schools: Correlation coefficients (raw and corrected for restriction in range)[1] between scores on six tests of the Basic Test Battery and final school grades for graduates of two types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test Form	Test: GCT r	GCT r_c	READ r	READ r_c	ARI r	ARI r_c	MAT r	MAT r_c	MKM r	MKM r_c	MKE r	MKE r_c
Gunner's Mates	16	2,183	1	.29	.39	.31	.42	.28	.36	.26	.33	.33	.38	.30	.38
	15	809	2	.19	.31	.22	.29	.27	.30	.33	.46	.41	.53	.38	.52
Torpedomen	10	1,251	1	.23	.33	.24	.33	.22	.29	.23	.28	.34	.39	.33	.40
	10	832	2	.13	.24	.16	.21	.21	.24	.25	.34	.36	.53	.26	.39

[1] Kelley, Truman L. *Statistical Method.* New York: Macmillan, 1924, p. 225.

TABLE 8-XII. Group 3 Schools: Means and standard deviations on six tests of the Basic Test Battery for graduates of two types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test Form	Test: GCT M	GCT σ	READ M	READ σ	ARI M	ARI σ	MAT M	MAT σ	MKM M	MKM σ	MKE M	MKE σ
Gunner's Mates	16	2,183	1	52.9	6.9	46.7	6.9	45.7	7.4	51.0	7.5	49.0	8.3	49.0	7.8
	15	809	2	59.3	5.4	61.3	7.2	58.2	9.0	58.6	6.7	56.8	7.0	56.9	6.7
Torpedomen	10	1,251	1	54.6	7.0	51.6	7.0	50.9	7.5	53.5	7.9	53.8	8.5	52.4	8.2
	10	832	2	62.5	5.1	64.9	7.4	62.3	8.9	61.7	7.1	63.8	6.0	62.2	6.2

gunner's mate and torpedoman trainees without the addition of other test scores.

SPECIAL TESTS OF THE BASIC TEST BATTERY. Three other tests are essentially a part of the Basic Test Battery but are treated separately here. They are the Spelling Test, Clerical Aptitude Test, and Radio Code Test—Speed of Response. The Radio Code Test and Form 1

TABLE 9-XII. Group 3 Schools: Correlation coefficients between sums of scores on two tests of the Basic Test Battery and final school grades for graduates of two types of elementary Naval Training Schools

Naval Training School	Test Form	Pair of Tests Whose Summed Score Has Highest Validity	Validity Coefficient Based on Summed Score	Multiple R	Validity Coefficients for MKM + MKE
Gunner's Mates	1	READ + MKM	.47	.47	.40
	2	MKM + MKE or MAT	.56	.56	.56
Torpedomen	1	READ + MKM	.42	.42	.42
	2	MKM + MKE or MAT	.49	.53	.48

TABLE 10-XII. Correlation coefficients (raw and corrected for restriction in range) [1] between scores on Form 1 of the Special Tests of the Basic Test Battery and final grades for graduates of seven types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test: Spelling r	Spelling r_c	Clerical Aptitude r	Clerical Aptitude r_c	Radio Code r	Radio Code r_c
Quartermasters	12	424	.17	.17	.29	.42		
Radio	16	1,013	.19	.22	.18	.25	.29	.39
Signal	8	635	.20	.23	.32	.46	.22	.26
Storekeepers [2]	4	363	.25	.25	.32	.43	.01	.01
Storekeepers (WR)	2	241	.28	.27	.33	.48		
Yeomen [3]	6	274	.27	.31	.39	.54	.28	.29
Yeomen (WR)	6	1,190	.35	.31	.33	.46		

[1] Kelley, Truman L. *Statistical Method.* New York: Macmillan, 1924, p. 225.
[2] 1 class, 64 graduates.
[3] 2 classes, 97 graduates.

of the Spelling Test and Clerical Aptitude Test were given over the same period of time as Form 2 of the tests previously discussed, and the validity data were gathered from the same classes for which Form 2 data have already been presented.

The validity coefficients for these three tests are shown in Table 10-XII and the corresponding means and standard deviations in

Table 11-XII. The Spelling Test does not predict success in any of the schools studied well enough to justify its addition to the other tests. The Clerical Aptitude Test has fairly good validity coefficients for most of the schools listed, and in the classes where comparison is possible it is the best single test for yeoman and signal schools.

TABLE 11-XII. Means and standard deviations on Form 1 of the Special Tests of the Basic Test Battery for graduates in seven types of elementary Naval Training Schools

Naval Training School	Number of Classes	Number of Graduates	Test: Spelling M	Spelling σ	Clerical Aptitude M	Clerical Aptitude σ	Radio Code M	Radio Code σ
Quartermasters	12		61.8	10.0	62.0	6.6		
Radio	16	1,013	59.8	8.5	60.5	7.2	68.1	7.1
Signal	8	635	61.6	8.6	60.9	6.6	61.7	8.3
Storekeepers [1]	4	363	61.6	10.0	63.1	7.1	59.1	9.9
Storekeepers (WR)	2	241	64.7	10.2	64.9	6.5		
Yeomen [2]	6	274	62.3	8.5	63.5	6.6	62.5	9.5
Yeomen (WR)	6	1,190	63.4	11.4	64.1	6.7		

[1] 1 class, 64 trainees.
[2] 2 classes, 97 trainees.

TABLE 12-XII. Comparative validity of single tests and combinations of tests of the Basic Test Battery in six types of elementary Naval Training Schools

Naval Training School	Correlation Coefficients between Test Scores and Final School Grades: Single Test GCT	ARI	CLER	Combination of Tests GCT + ARI	GCT + CLER	GCT + ARI + CLER
Quartermasters	.48	.41	.42	.47	.49	.49
Signal	.44	.37	.46	.43	.49	.48
Storekeepers (WR)	.52	.44	.48	.51	.55	.54
Storekeepers	.58	.54	.43	.59	.55	.58
Yeomen	.47	.35	.54	.43	.55	.51
Yeomen (WR)	.65	.50	.46	.61	.61	.60

The Radio Code Test is effective only in the case of radio schools and is the only one of the nine tests that appears to be useful for that school.

Table 12-XII gives information on the comparative validity of the General Classification Test, Arithmetical Reasoning Test, and

Clerical Aptitude Test, and the sums of the most plausible combinations. In these six types of schools the best single test appears to be about as satisfactory as the best combination of tests in predicting success.

VALIDITY OF THE INTERVIEW. The effectiveness of the interview as a selection device is difficult to estimate because interviewers always had test scores before them as they made their evaluations. The quantitative result of the classification interview was recorded as the "quality classification" and is described earlier in this chapter.

A sample group of 37,862 trainees was studied to observe the relation between quality classification and success in nine types of elementary schools. The trainees were divided into three groups according to whether they (1) failed the school, (2) graduated as qualified strikers (apprentices for a rating), or (3) in the case of those

TABLE 13-XII. Distribution of trainees in nine types of elementary Naval Training Schools according to quality classification and success in training

Relative Success Categories	Cases assigned to Quality Classification by Classification Interviewers									
	1.		2.		3.		General Detail		Total	
	N	%[1]	N	%[1]	N	%[1]	N	%[1]	N	%[1]
Failure	178	3.47	636	4.11	1,482	9.70	507	25.68	2,803	7.40
Graduated as Qualified Striker	4,110	80.15	13,016	84.09	13,047	85.38	1,441	73.00	31,614	83.50
Graduated with Rating	840	16.38	1,827	11.80	752	4.92	26	1.32	3,445	9.10
Total	5,128		15,479		15,281		1,974		37,862	

[1] Base for percentage is total in the quality classification.

with pronounced proficiency, received petty officer rating at graduation. The contingency table showing the relationship between the two variables is shown in Table 13-XII. The coefficient of mean square contingency is .23. Although this denotes some relationship, it is not as high as the validity coefficients of the appropriate tests for the same schools. The fact that the data for all classes of all nine schools were pooled may serve to obscure somewhat the predictive efficiency of the quality classification.

The four categories of quality classification evidently contribute unequally to the relationships; that is, there is comparatively little difference between classifications 1 and 2, and only a moderate difference between classifications 3 and 1 or classifications 3 and 2. The sharp difference is between "general detail" on the one hand and classifications 1, 2, and 3 together on the other. In other words,

the interviewer was only moderately successful in predicting who would do best among those recommended for school training, but he was markedly successful in determining who should go to school and who should go to "general detail".

The actual order of assignment to school involved both quality classification and the school recommendation. The interviewer might assign as a first recommendation either (1) a recommendation for a particular elementary school, or (2) a recommendation to general detail. If the first recommendation were for assignment to an elementary school, it was mandatory that the second recommendation also be to an elementary school. If the first recommendation were assignment to general detail, the second recommendation might be assignment to an operational school. When both quality classification and recommendation were considered, the contingency coefficients between order of assignment and graduation ranged from .22 to .35. They show throughout a lower degree of relationship than the appropriate tests for the same schools.

In another study of 3,246 graduates of electrical schools, it was found that the correlation ratio between quality classification and school grades was .41. For the same group, the uncorrected correlation coefficient between school grade and the sum of the Arithmetical Reasoning Test plus the Mechanical Knowledge Test (Electrical Score) was .50. To evaluate concretely this difference in predictive efficiency, the success of quality classification 1 men was compared with that of the same number of top scoring men on the Mechanical Knowledge Test (Electrical Score) alone. Of the 333 men in quality classification 1, ten failed and eighty-one received a rating at graduation. Of the top 333 men in the Mechanical Knowledge Test (Electrical Score) distribution, four failed and one hundred were rated. It should be noted, however, that it is not always true that a quality classification of 1 denotes an interviewer's judgment of superiority over quality classification 2. Actually a man may be given a 1 because he is *disqualified* for all but one type of school, while a 2 man may not only be better qualified for that particular school, but is also qualified for several others.

Both studies reported here on the use of quality classification indicate that the improvement in predicting school success by having in addition to test scores an interviewer's evaluation of experience, interest, and personality, is relatively small and may well be negative. In this connection it should be noted that prediction of school success is but one of the functions of the interview. The Enlisted Personnel Qualifications Card is most conveniently filled out in the classification interview, and it presents an opportunity for explaining the nature of the assignment to the interviewee and answering

TABLE 14-XII. Distribution of trainees in eight types of elementary Naval Training Schools according to age

Naval Training School	Age Group	Percentage of Trainees in Each Group	Final School Grade	
			M	σ
Aviation Machinist's Mates	17-21	33.3	82.15	3.69
(N = 766)	22-26	44.3	84.34	2.61
	27-30	13.3	84.19	3.33
	31-38	9.5	83.51	3.47
	All AMM Trainees		83.52	3.33
Aviation Ordnancemen	17-21	49.0	84.48	3.77
(N = 1,230)	22-26	25.4	86.31	3.93
	27-30	15.6	86.08	3.60
	31-38	10.0	86.08	4.21
	All AOM Trainees		85.36	3.93
Aviation Radiomen	17-19	70.2	84.45	4.56
(N = 1,125)	20-22	17.1	85.04	4.86
	23-25	8.7	84.29	4.12
	26-38	4.0	84.68	4.67
	All ARM Trainees		84.55	4.58
Electrical (N = 1,930)	17-21	27.6	82.90	7.02
	22-26	22.3	85.06	6.74
	27-31	23.2	84.95	6.62
	32-36	21.8	84.43	6.88
	37-42	5.1	83.95	7.13
	All EM Trainees		84.24	6.90
Quartermasters (N = 1,189)	17-21	25.1	85.67	7.59
	22-26	31.9	85.58	7.13
	27-31	21.7	85.55	6.08
	32-36	17.8	85.92	6.88
	37-44	3.5	85.02	5.50
	All QM Trainees		85.64	6.94
Radar Operators	17-21	10.5	81.07	6.45
(N = 1,946)	22-26	41.8	82.39	6.47
	27-31	23.9	82.06	6.33
	32-36	18.8	81.57	6.61
	37-42	4.9	79.21	6.79
	All RdM Trainees		81.86	6.52
Radio (N = 1,153)	17-21	60.7	87.63	6.17
	22-26	18.5	87.02	6.36
	27-31	12.9	86.81	5.75
	32-38	7.9	86.24	6.73
	All RM Trainees		87.30	6.21
Signal (N = 1,433)	17-19	67.4	83.99	6.87
	20-23	10.6	84.25	7.26
	24-27	10.5	86.00	6.47
	28-31	7.0	87.08	6.00
	32-38	4.5	85.73	7.25
	All SM Trainees		84.52	6.90

any questions he may have. The interviewee also has an opportunity to express his interests and point of view, and quite apart from any predictive value the information may have, the mere fact of his having had individual attention and an opportunity to be heard is probably valuable from the morale point of view.

AGE AS A PREDICTIVE FACTOR. The relation between age and school grades is shown for several schools in Table 14-XII. In general, men between 22 and 36 years of age are more successful than those below 22 or over 36.

The product-moment correlation coefficients between age and grades assigned in school are shown in Table 15-XII. Since the relationship has been found to demonstrate a pronounced curvilinearity, the correlations understate the amount of relationship present. For

TABLE 15-XII. Correlation coefficients (product-moment and biserial) between age and final grade for trainees in eight types of elementary Naval Training Schools

Naval Training School	Product-moment r	Biserial r[1]
Aviation Machinist's Mates	.200	.379
Aviation Ordnancemen	.195	.275
Aviation Radiomen	.008	.040
Electrical	.081	.155
Quartermasters	.002	.006
Radar Operators	.081	.135
Radio	—.069	—.080
Signal	.141	.143

[1] The dichotomy was between "favored age" and "non-favored age," where the favored age group was between 22 and 36 years of age and the non-favored ages were below 22 and above 36.

classification purposes it is sufficient to divide the age range into a favored range, 22-36, and unfavored ranges, 21 and below, and 37 and above. The amount of predictive efficiency that could be achieved by this practice can be evaluated by a biserial coefficient of correlation, even though it is based on a derived variable, "favored", rather than on the age scale itself. The only such biserial coefficients indicating a useful relationship for prediction are found for aviation machinist's mate school, .40, and aviation ordnanceman school, .28.

AMOUNT OF EDUCATION AS A PREDICTIVE FACTOR. The relationship between amount of civilian education in years and school grades is shown in Table 16-XII. As in the case of age, the regression appears curvilinear. In Table 17-XII are shown the product-moment correla-

TABLE 16-XII. Distribution of trainees in eight types of elementary Naval Training Schools according to amount of civilian education in years

Naval Training School	Amount of Civilian Education in Years	Percentage of Trainees in Each Group	Final School Grade	
			M	σ
Aviation Machinist's Mates (N = 766)	0-8	3.8	82.64	2.97
	9, 10, 11	27.4	82.09	3.62
	12	57.8	84.00	3.11
	13 and over	11.0	84.86	2.43
	All AMM Trainees		83.52	3.33
Aviation Ordnancemen (N = 1,230)	0-8	6.4	84.17	4.08
	9, 10, 11	35.0	84.16	3.79
	12	50.1	86.12	3.76
	13, 14, 15	8.0	86.76	3.77
	16 and over	0.6	84.78	3.28
	All AOM Trainees		85.36	3.93
Aviation Radiomen (N = 1,125)	0-8	1.9	84.31	3.49
	9, 10, 11	32.3	83.83	4.53
	12	58.0	84.83	4.62
	13 and over	7.7	85.46	4.37
	All ARM Trainees		84.55	4.58
Electrical (N = 1,930)	0-8	11.3	81.72	6.89
	9, 10, 11	34.9	82.15	7.01
	12	46.7	85.90	6.27
	13, 14, 15	6.3	87.79	5.75
	16 and over	0.9	86.38	5.92
	All EM Trainees		84.24	6.90
Quartermasters (N = 1,189)	0-9	2.3	83.90	6.45
	10, 11	7.7	83.72	7.08
	12	51.4	85.57	7.12
	13, 14, 15	29.4	86.24	6.32
	16 and over	9.3	86.15	7.28
	All QM Trainees		85.64	6.94
Radar Operators (N = 1,946)	0-8	1.6	80.32	7.24
	9, 10, 11	14.8	79.40	6.69
	12	62.0	82.02	6.26
	13, 14, 15	16.6	82.72	6.77
	16 and over	5.1	84.62	5.89
	All RdM Trainees		81.86	6.52
Radio (N = 1,153)	0-8	6.4	87.68	5.97
	9, 10, 11	33.1	85.86	6.37
	12	51.2	88.08	5.98
	13, 14, 15	8.6	87.97	6.31
	16 and over	0.7	87.25	4.79
	All RM Trainees		87.30	6.22
Signal (N = 1,433)	0-8	5.7	82.96	7.27
	9, 10, 11	41.7	82.81	6.64
	12	44.8	85.57	6.76
	13 and over	7.8	88.69	5.69
	All SM Trainees		84.52	6.90

tion coefficients and the biserial coefficients. For the latter the dichotomy was completion of high school versus non-completion. Useful relationships are found in the following cases:

	$r_{bis.}$
Electrical school	.37
Aviation machinist's mate school	.36
Aviation ordnanceman school	.32
Radar operator school	.24

Before these data are used as a basis for assignment recommendations, it would be necessary to know the relationship of test scores to both education and grades for these graduates, so that the predictive efficiency of years of civilian education, with appropriate test scores partialed out, could be determined. If the correlation coefficients between years of education completed and the tests of the Basic Test Battery reported in Chapter VI are applied to this group, the correlation coefficients between years of education and service school grades, with test scores held constant, are very close to zero.

TABLE 17-XII. Correlation coefficients (product-moment and biserial) between amount of civilian education in years and final grades for trainees in eight types of elementary Naval Training Schools

Naval Training School	Product-moment r	Biserial r[1]
Aviation Machinist's Mates	.174	.360
Aviation Ordnancemen	.156	.324
Aviation Radiomen	.061	.181
Electrical	.219	.371
Quartermasters	.049	.156
Radar Operators	.15	.244
Radio	.008	.188
Signal	.161	.156

[1] The dichotomy was between less than 12 years of education and 12 years or more.

It should be noted, however, that the trainees for these schools had been selected on the basis of test scores, which in turn restricted the variability in years of education and lowered its apparent predictive value. In other words, if large numbers of men with less than eight years of civilian education had been sent to elementary enlisted service schools, their high rate of failure would have emphasized the importance of civilian education. In general, it is probable that "years of education" contributes little in addition to test scores, but that it is a useful substitute for tests if the latter are not available.

GENERAL CLASSIFICATION TEST SCORES IN RELATION TO REASONS FOR FAILURE. The records of 46,500 trainees assigned to ten different types of elementary schools were examined to determine the relationship between scores on the General Classification Test and failure in school. Of the trainees, 11,000 obtained General Classification Test scores which were below the cutting score recommended for the school to which they were assigned. Presumably such assignments had been made because of quota pressure and because those individuals had some such unusual qualifications as specific experience, marked interest, or favorable personal qualities, which might

TABLE 18-XII. Percentage of failure among trainees of ten types of elementary Naval Training Schools whose scores on the General Classification Test were (1) below the cutting score; and (2) above the cutting score, classified according to reasons for failure

Reason for Failure	Percentage of Failure Among Men with GCT Scores Below Cutting Score. N = 11,000		Percentage of Failure Among Men with GCT Scores Above Cutting Score. N = 35,500	
Lack of Aptitude				
In general aptitude for school	5.5		1.2	
For practical work	3.0		0.8	
Total		8.5		2.0
Poor Motivation				
Lack of interest or appreciation	4.9		1.2	
Preferred sea duty or other school	.5		.3	
Total		5.4		1.5
Disciplinary Problem		.8		.3
Miscellaneous (death, illness, discharge, etc.)		1.1		1.1
All Reasons		15.8		4.9

counterbalance a slight deficiency in the General Classification Test score. The scores on other tests of the Basic Test Battery were not taken into account in this study. Table 18-XII shows the percentage of the "above cutting score" and "below cutting score" groups who failed for each reason, for reasons related to inaptitude or poor motivation, and for all reasons. Despite their non-test qualifications, the rate of failure of men with scores below the cutting score was more than three times that of men who met the cutting score. The contrast becomes sharper if the comparison is made only on factors for which prediction by selection tests can be expected. If the miscellaneous reasons such as death, discharge, and serious illness are

excluded, it is found that almost four times as many of those below cutting score fail as those above cutting score. This represents a degree of predictive efficiency equivalent to a tetrachoric coefficient of correlation of .53.

The reasons for failure presented in Table 18-XII are those given by the school instructors. While the reasons as given are probably of doubtful validity, the data for the reason "lack of interest or application" are of special interest. More than four times as many of the below cutting score group failed for lack of interest as did those who were above the cutting score. The predictive efficiency is equivalent to a tetrachoric coefficient of correlation of .51. This is precisely the area of selection for which an aptitude test might be considered least useful and for which subjective and intuitive interview data are considered necessary, and yet rigid mechanical use of objective test scores would have reduced attrition from alleged poor motivation very substantially. When it is considered that many of the below cutting score groups were presumably assigned just because their high interest and serious intent were judged to compensate for their low aptitude, it is apparent that the entire area of "interest," as it appears to the classification interviewer and to the school instructor, merits further study. It may be, for example, that "lack of interest" is the result of lack of aptitude, or it may be that "lack of interest" is a charitable way of describing a cause of school failure. In any case the whole subject requires much more intensive research.

Summary of Results by Predictive Factors

The Basic Test Battery—General Tests

Verbal Tests. The General Classification Test and the Reading Test predict success well in most of the schools studied. Only in the radio school and in three schools with a distinct mechanical emphasis (machinist's mate, gunner's mate, and torpedoman) do the corrected validity coefficients fall consistently below .40. The two tests are similar in predictive efficiency, with the General Classification Test somewhat more effective, especially in Form 2. For purposes of prediction, little justification is found for using a test of reading ability in addition to the General Classification Test for the measurement of verbal ability.

Arithmetical Reasoning Test. The data from Form 1 indicate that the Arithmetical Reasoning Test is the most generally useful test of the battery, considering all schools in all groups. This conclusion is substantiated by the results presented in the fleet validation study

(Chapter XX) and by data from a research study carried out aboard the *USS New Jersey* showing a close relationship between scores on the Arithmetical Reasoning Test and attainment of petty officer status. The kind of ability measured by this test appears to be generally useful in naval ratings even when the school or the job is not notably quantitative in nature. The lower validity coefficients for the Arithmetical Reasoning Test on Form 2 may be partly attributable to the very high mean scores of the classes studied, with the result that a large portion of each class had as high a level of the ability as the school demanded. This explanation implies a curvilinear relationship between school grades and scores on the Arithmetical Reasoning Test.

The Mechanical Tests. These three tests (the Mechanical Aptitude Test, and the Mechanical and Electrical Scores on the Mechanical Knowledge Test) predict success in the schools with an obvious mechanical emphasis. But the data provide little basis for assigning personnel to different mechanical schools by their different scores on the three tests. The Electrical Score of the Mechanical Knowledge Test, for example, is as good as the Mechanical Score and better than the Mechanical Aptitude Test score for predicting success in the diesel, machinist's mate, and gunner's mate schools. This would appear to indicate that electrical knowledge is highly important in schools where, in fact, electrical subjects are not stressed.

A possible explanation of the anomaly is that the Mechanical Knowledge Test (Electrical Score) is operating as a test of mechanical aptitude. Many of the items rest upon fairly casual observation of common electrical equipment such as toasters, telephones, and doorbells. A large proportion of the testees have had the necessary opportunity and environmental exposure to arrive at the answer if they have the requisite aptitude and interest. On the other hand, many of the Mechanical Knowledge Test (Mechanical Score) items relate to specific tools which are encountered only in a metal working shop. The fact that more people have had the necessary experiences for answering the Mechanical Knowledge Test (Electrical Score) items correctly tends to make this test operate as an aptitude test. The aptitude involved apparently is sufficiently general to be important in clearly mechanical schools.

The Mechanical Aptitude Test has two sections testing (1) spatial perception and (2) mechanical comprehension; it may, therefore, be on too abstract a level to correlate highly with success in these mechanical schools. In other words, a test constructed and designated as an electrical knowledge test may measure certain aspects of mechanical aptitude better than tests designed to measure mechanical knowledge or mechanical aptitude. In this connection it may be

noted that in Forms 4 and 5 of the Basic Test Battery, mechanical comprehension, mechanical knowledge, and electrical knowledge items have been combined to form the Mechanical Test.

The Special Tests

Spelling. The Spelling Test showed insufficient predictive efficiency to justify its continued use, especially in view of its high intercorrelation with the General Classification Test.

Clerical Aptitude Test. The Clerical Aptitude Test predicts success well not only in the yeoman and storekeeper schools, but also in quartermaster and signal schools. Insufficient data are available to evaluate the test's usefulness for a wider range of schools, but the validity coefficients for quartermaster and signal schools (where the clerical aspect is not especially prominent) suggest that the Clerical Aptitude Test may measure speed and accuracy of repetitive perceptual operations.

Radio Code Test—Speed of Response. This test predicts success fairly well in the radio school only, and it is the only test that is effective for radio school. It should be noted, however, that the data in Chapter XX show a substantial correlation between the General Classification Test and success of radiomen in the fleet.

Since these conclusions on the nature of the functions involved in the tests of the Basic Test Battery are somewhat speculative, considerable caution is indicated in setting up selection requirements on the basis of the names of the tests and the obvious features of the assignment. Since an electrical knowledge test predicts success in schools which are not electrical, an arithmetical reasoning test is generally useful for most schools whether or not they stress arithmetic, and a clerical aptitude test works well for schools that are not particularly clerical in content, the need for caution in using tests because they have the "right" names is indicated.

Other Predictive Factors

Interview. Data on the "quality classification" indicate that the interview adds little if anything to the predictive value of the tests alone, and that some reasons for school failure, such as "lack of interest or application," can be predicted by aptitude tests, although they are usually considered as factors which should be appraised by the interview technique.

Age. Trainees between 22 and 36 years of age receive somewhat better school grades than those beyond those limits, but the relationship is not close.

Education. Amount of civilian education, in the absence of aptitude test scores, is of some value in predicting school success but

contributes little additional information when test scores are available.

The data presented in this chapter indicate that success in elementary enlisted schools can be predicted effectively by means of the general and special tests of the Basic Test Battery. The most effective combination of tests for Group 1 schools is that of the General Classification Test and the Arithmetical Reasoning Test. Success in Group 2 schools is best predicted by the Arithmetical Reasoning Test plus the Mechanical Knowledge Test (Electrical or Mechanical Score), depending upon the particular type of school. The Mechanical and Electrical Scores of the Mechanical Knowledge Test proved the best combination for Group 3 schools. Judging from the results obtained from Forms 1, 2, and 3 of the Basic Test Battery, the four tests (singly and in combination) of Forms 4 and 5, as described in Chapter VI, should provide effective tools for the prediction of success in the Navy's elementary enlisted schools.

CHAPTER XIII

PREDICTION OF SUCCESS IN ADVANCED SERVICE SCHOOLS

ADVANCED enlisted service schools are designed to provide instruction at advanced levels for men who have had elementary school training or experience in rating, ashore or afloat (Chapter V). Until nearly the end of the war, trainees were assigned to these schools without the benefit of systematic selection requirements. Owing to pressure of classification and training problems on the elementary level, studies in the prediction of success in advanced schools were postponed until late in the war. Two studies were begun in the summer and early fall of 1944 in (1) Naval Training School (Fire Control—Advanced), and (2) Naval Training School (Gunner's Mates and Electric Hydraulics). Preliminary reports on these studies, recommending cutting scores for selection of trainees, were issued in December 1944 and January 1945. The rest of the studies summarized in this chapter were not begun until April 1945. At that time the Test and Research Section was requested by the Enlisted Classification Section to initiate studies in certain advanced schools leading to the preparation of selection requirements comparable to those which had been worked out at the elementary school level.

Nature of Advanced School Programs

The researches reported in this chapter are concerned with only a fraction of the total number of advanced enlisted schools for naval personnel. All of the schools studied were concerned with qualifying enlisted personnel for Navy jobs requiring skill in operation of equipment, in repair and maintenance of equipment, or in both. Notwithstanding this superficial similarity in type of training, the schools varied widely with respect to (1) length of training program, (2) type of content, (3) amount of pre-selection of trainees, and (4) standards of achievement. A list of the schools studied, together with the length of training programs, and the general nature of the curricula is given below.

ADVANCED SCHOOLS FOR MOTOR MACHINIST'S MATES

1. NTSch (Diesel)-A. 18 weeks. Advanced general training in diesel engines—not a factory school concerned with a special make of engine.

2. NTSch (Diesel)-B. 4 weeks. A factory school providing specialized training on certain types of diesel engines.

3. NTSch (Diesel)-C. 5 weeks. A factory school.

4. NTSch (Packard Marine Engine). 8 weeks. A factory school.

Advanced Schools for Electrician's Mates

5. NTSch (Electrical Interior Communications) (EIC School). 26 weeks. The program of this school is highly intensive and extremely technical; the most advanced of any of the electrician's mates schools. The curriculum contains a great deal of mathematics (including some calculus) and electronics theory at the level of the last two years of an electrical engineering course. Trainees for this school were selected from the schools described below.

6. NTSch (Electrician's Mates). 18 weeks. An advanced electrician's mates school. The curriculum includes considerable mathematics and electronics theory, but the level is less advanced than that of the EIC School.

7. NTSch (Gyro Compass). 16 weeks. An advanced electrician's mates school. Specialized training is provided in repair and maintenance of gyro circuits and equipment. The level of training is somewhat advanced compared to school 6, but not so intensive as school 5.

Advanced School for Gunner's Mates

8. NTSch (Gunner's Mates and Electric Hydraulics). Length of course varied—average 12 weeks. The curriculum covers basic electricity and fire control, basic hydraulics, 40 mm. gun and mount, and 5″/38 caliber gun and mount. This school was included in the set of studies begun in April 1945, although it had been studied earlier, in the summer and fall of 1944. By the time the later set of studies was begun, two curricula were being offered. These differed mainly in level of difficulty. The potentially better students were placed in an advanced course; the remaining students in a primary course. Promising graduates of the primary course sometimes were placed in the advanced course.

Advanced School for Fire Controlmen

9. NTSch (Fire Control—Advanced). Length of course variable—consisted of a basic course of 8 weeks, and specialized courses of varying lengths (3 weeks to 24 weeks) concerned with special types of equipment. Most of the specialized courses were 8 or 10 weeks in length. A preparatory course of 4 weeks was given for men considered deficient in background for the basic course. The basic

course covered elementary electricity, basic fire control, basic hydraulics, and standard assemblies. The specialized courses were concerned with gun directors, computors, and other fire control equipment for particular guns. The longest course (24 weeks) was a repair course covering a wide variety of fire control equipment.

ADVANCED SCHOOL FOR WATER TENDERS

10. NTSch (Oil Burning). 6 weeks. Advanced training in operation and maintenance of high pressure steam boiler systems.

ADVANCED SCHOOLS FOR SONARMEN

11. West Coast Sound School. 5 weeks. Advanced training in operation and maintenance of sonar equipment and in plotting and interpreting information received through sonar.

12. Fleet Sonar School. 6 weeks. Substantially the same as West Coast Sound School.

The above listing indicates marked variations in training time, type of content, and level of difficulty of different advanced school programs. No less disparity was found with respect to achievement standards. The EIC School, for example, applied high and rigid standards. Its failure rate in the classes studied was from twenty to twenty-five per cent. On the other hand, the samples for most schools included very few failures, in some instances none. In some schools the practice of permitting students to fail was contrary to policy. Reasons advanced for this policy, despite the consequent lowering of standards, were: (1) failure could not be afforded because of the Navy's need for men with some degree of training in the particular skills concerned, (2) trainees were sometimes rated at pay grade levels above those for which they were actually qualified, and the school's job was to supply the missing knowledge or skill, no matter how resistive the trainee might be to the process, and (3) to have men repeat portions, or all, of the curriculum until some degree of mastery was attained was considered to be better policy than to fail them.

Schools also differed with respect to the degree of selection of their trainees. Discussion of this point will be reserved for a later section.

Populations Available for Selection of Trainees

Trainees for the advanced service schools came from a number of different sources. Some were graduates of elementary schools who

were immediately sent on for advanced training; some were assigned from shore duty stations to prepare them for further duty ashore or afloat; still others were men returned from sea duty. All varieties and lengths of Navy experience were represented. Not infrequently a given school group included all pay grades from unrated men to chief petty officers. The age distribution was correspondingly wide.

The heterogeneity of Navy experience represented in the advanced school population complicated the prediction problem. The relationships between experience and such factors as motivation, adjustment to the school situation, study habits, and personal and family problems, are difficult to evaluate but are nevertheless real in their influence on training success. It seems reasonable to expect less uniformity, with respect to these imponderables, in the populations with which these studies dealt than in the groups assigned to elementary schools. Training officers frequently expressed the opinion that men returned from sea duty were poorly motivated. They tended to regard their training assignment as a shore duty vacation which had been earned by their months or years in the fleet. Reports of training officers indicated, also, that men returned from sea duty had more personal and family problems that interfered with study, and that in general they experienced more difficulty in adapting to the school situation.

The variations in composition of the groups available for selection of advanced school trainees at different times may be illustrated by data from the advanced fire control school. For four groups studied, representing four different enrollment times, the ranges in proportion of men from various sources were as follows: (1) returned from sea duty, 41 per cent to 85 per cent; (2) from elementary service school, none to 55 per cent; (3) from instructional duty (generally at elementary service schools), from none to 13 per cent; (4) from other duty, 1 per cent to 13 per cent. As time went on, more and more men returning from sea duty became available for advanced training and fewer elementary school graduates were immediately assigned to advanced training. But within this long time trend the composition of the population available for assignment to any school was also subject to short time fluctuations. The composition of these populations, at any given time, was therefore a variable and uncertain matter.

Procedures

SAMPLES STUDIED. The foregoing discussion of the populations available for selection of advanced school trainees indicates that the problem of adequate sampling was difficult. Solution of it would have required careful preliminary study of the groups available for

assignment to each school, followed by design of sampling procedures that would insure adequate representation to each subgroup. Where subgroups were too unlike to warrant combining them into a total sample, sampling within the subgroups should have been extensive enough to permit separate analyses of predictive relationships for each group.

Wartime pressure for immediate results precluded working out and employing such carefully designed sampling procedures. Samples were composed of intact school classes. For the most part the only requirement imposed on the samples was inclusion of a certain minimum number of cases. Even this requirement was not met in a few cases.

Only by virtue of some fortunate sampling accident would samples drawn in this way be representative of the total population available for selection of trainees. Time periods covered by the samples for most schools were too short to insure adequate representation of all sources from which men might be assigned. Numbers of cases were too small. Departure from representative sampling was, indeed, almost insured by whatever selection procedures were employed in choosing men for school assignment and because certain groups of cases had to be largely discarded for lack of data on them.

Although there were no systematic established selection standards for advanced schools, selection processes of various sorts were operating. Probably the most systematically selected groups were those assigned from elementary schools. The elementary school graduates selected for immediate assignment to advanced school were usually those who had proven most apt in their elementary training. Trainees for NTSch (Electrical Interior Communications) were a special case of this kind of selection, although in this instance selection was made from trainees assigned to another advanced school, rather than to an elementary school. The next most systematically selected groups were those recommended by advanced classification centers. Lack of established selection standards probably prevented these groups from being more systematically assigned. One would expect locally determined standards, based on intuition rather than research, to vary from classification center to classification center. Moreover, aptitude was not the only basis for assignment. Classification centers were directed to consider, also, the extent to which men needed advanced training in order to qualify for advancement, or in some cases, to supply missing qualifications for ratings already held. Probably the least systematically selected groups were those recommended by commanding officers or personnel officers of previous duty stations. The suspicion has even been expressed by some training officers that commanding officers, at least occasionally,

tended to use the advanced schools as dumping grounds for their undesirables.

A considerable number of cases had to be eliminated from some samples because data on them were lacking. In only two of the studies, at NTSch (Gunner's Mates and Electric Hydraulics) and at the Advanced Fire Control School, were the tests given at the schools by the investigators. In all other studies the schools were requested to supply the data from the men's Enlisted Personnel Qualifications Cards. Men who had come into the Navy prior to the establishment of classification centers in recruit training stations often had no Qualifications Cards in their service record jackets. Occasionally such men had been interviewed and had a Qualifications Card partially filled out, but frequently so much data was lacking that the cards were of little value for research purposes. A large proportion of men with long sea experience fell in this group. Hence, the samples were systematically biased by the exclusion of individuals for whom data were not readily available.

PREDICTION DATA. The prediction data for these studies were confined almost entirely to those conveniently available from the Enlisted Personnel Qualifications Cards. These included Basic Test Battery scores, rating and pay grade, age, Navy training records, length of Navy service, civilian occupation, and civilian education. In only one instance (the Electrical Interior Communications School) were other prediction data used. In this case, since all men were assigned to this school from NTSch (Electrician's Mates), their previous school grades were obtained.

Certain limitations of these Qualifications Card data should be noted.

1. For some schools, data with respect to certain prediction variables, particularly scores on certain tests, were too incomplete to permit estimation of the relationships of those variables to success in the schools concerned.

2. The data on civilian occupational experience was unsatisfactory. It was necessary to rely on the codes of civilian occupations that had been recorded by classification interviewers. Even assuming that coding was always accurate, the codes had the following flaws: (a) Different coding systems were used in the classification program at different times, and the particular coding system employed on each individual could not always be determined. (b) The same numerical code was sometimes employed in two coding systems, but with different meanings. (c) One of the coding systems failed to provide a means of making important distinctions concerning degree of skill in certain occupational groups. Data not coded in conformance with the Dictionary of Occupational Titles were discarded.

3. Men took tests at widely different stages of Navy experience, a fact which probably influenced their scores on certain of the tests. Those men who were inducted into the Navy after classification centers were established in recruit training stations were tested during their recruit training. Men who entered the Navy prior to the establishment of the classification program were tested, if at all, following months or years of Navy experience. For such tests as the Mechanical Knowledge Test (both Mechanical and Electrical Scores) this difference in stage of Navy experience at which men were tested was probably important. Testing a man's knowledge of tool relationships after he has had several years of experience as a motor machinist's mate does not measure aptitude for diesel school training in the same way that it is measured if the test is taken prior to such experience.

4. The Basic Test Battery was designed primarily as a tool for classification at the recruit training level to aid in assignment of men to training in elementary schools. It was not intended for the special job of measuring aptitudes for advanced school training. It is probable that better prediction at the advanced school level could be obtained from tests designed especially for the purpose.

5. Insufficient data were available on Navy experience. Data on this factor were limited to training records, length of Navy experience, rating, and pay grade. No account was taken of duty assignments, except as they were related to the men's ratings and pay grades. Yet particular duty assignments within their ratings probably determined in considerable measure the backgrounds of knowledge with which men began advanced training.

Criterion Data. The criterion of success used was final grades, or rank in class based on final grades. Whenever several classes were combined into a total sample, grades were converted to ranks. Use of ranks for combining classes assumes equivalence, among classes, of true criterion score means and standard deviations. This assumption was probably not well satisfied, particularly where classes were small. But it was considered preferable to the alternative assumption that grading standards were constant from class to class. Studies in elementary service schools have shown that class to class variations in grading standards may be of considerable magnitude. Because differences in class size prohibited combining ranks directly, such ranks were converted to linear scale values, assuming a normal distribution of true criterion values.

The acceptability of these grades as a criterion was not a question in the design of the studies. No other criterion was available, and time could not be taken to develop new criterion measures or to refine existing grading systems. From one point of view, the grades

assigned by the schools are, by definition, the criterion of success. Certainly the immediate concern of the Navy was to assign to schools those trainees who could pass the courses. It was necessary to reduce attrition rates as much as possible, so that the largest number of trained men could be available to the fleet in the shortest period of time. Except for a program designed to meet an immediate urgent need, however, the question of how accurately the grading systems reflect the ability of school graduates to perform on the job ought not to be ignored. Studies are badly needed to validate both the curricula and the grading systems against job performance.

Final grades were generally averages of quizzes and written examinations together with laboratory performance ratings of some sort. Paper and pencil quizzes and examinations in the advanced schools were ordinarily unstandardized and of unknown reliability and validity. They were usually made up by the instructional staffs of each school, most of whom were not skilled or trained in test construction. Such tests quite frequently measure a verbal factor more than the knowledge and skills which they are designed to measure. They may, therefore, do an injustice to the student who lacks verbal facility, but who, nevertheless, has mastered the essentials of the curriculum.

Laboratory grades were generally based on instructors' observations of the very limited number of men under their immediate laboratory supervision, and the grades were assigned without the aid of standardized laboratory performance tests or carefully constructed check lists or rating scales. Moreover, instructors were ordinarily not skilled raters, and were as subject to the usual "central tendency" and "leniency" errors as any other unskilled raters. Statistical evidence on this point was available from only one school. School records, in this instance, made it possible to obtain separate averages for laboratory grades and written test grades, as well as a final average including both. This school seems to have been more than usually aware of the difficulties involved in assigning such ratings and had made efforts in the direction of improving and standardizing rating procedures. Even so, the dispersion of laboratory marks was much smaller than that of quiz and examination grades. The average standard deviation of laboratory marks for two classes in this school was 2.4 as compared to an average standard deviation of written test grades of 6.7.

Part grades were usually combined into a final mark by taking the simple average of the part grades, or by computing a weighted average in which the nominal weights assigned were made proportional to the judged importance of the part grades. It is questionable whether functional weights were ever considered. In the case of the

TABLE 1-XIII. Means and standard deviations of available test scores and other selected personal data for trainee samples from twelve advanced Naval Training Schools for Enlisted Personnel

School	N		Test Score									Other Personal Data			
			GCT	ARI	MAT	MKM	MKE	READ	SPELL	CLER	RADIO	Mo. of Active Duty	Pay Grade[1]	Age in Years	Years of Civilian Education
For Motor Machinist's Mates															
Diesel A	148	M	51.6	49.7	52.0	56.4	55.0					27.9	3.5	24.5	10.6
		σ	8.1	9.9	8.4	7.5	8.3					9.8	.7	4.7	1.6
Diesel B	107	M	56.6	55.8	56.9	61.5	57.4					16.8	4.1	23.5	10.9
		σ	7.1	9.3	8.2	7.3	7.5					12.3	1.1	5.2	1.8
Diesel C	40	M	54.8	54.0	55.7	61.1	59.4					19.2	4.3	21.9	11.0
		σ	9.0	11.3	8.0	7.1	8.2					11.3	1.3	4.3	1.3
Packard Marine Engine	106	M	57.4	60.7	59.6	63.7	60.4		52.1	51.6	57.2			23.3	11.3
		σ	6.1	8.8	7.2	4.6	7.2		8.3	6.6	9.4			3.4	1.4
For Electrician's Mates															
Electric Interior Communications															
Group I	119	M	62.2	63.7	64.9	63.9	69.4	64.1				35.8	3.1	26.4	12.0
		σ	6.2	8.7	7.2	7.0	5.9	8.4				22.9	1.2	4.9	1.2
Group II	109	M	61.9	63.7	63.1	64.9	70.4					36.4	2.8	24.6	11.9
		σ	6.8	7.8	7.3	5.3	4.0					19.1	1.2	3.8	1.2
Electrician's Mates	205	M	55.2	52.8	55.4	60.4	66.2					38.5	3.0	25.9	10.8
		σ	7.3	9.7	8.3	7.7	6.3					15.4	.9	4.9	1.5
Gyro Compass	59	M	55.2	54.7	54.7	59.5	64.8					36.8	3.3	23.5	10.8
		σ	7.3	9.9	8.9	7.4	8.0					14.5	.8	3.8	1.6
For Gunner's Mates															
Gunner's Mates and Electric Hydraulics															
1st Study	154	M	53.9	50.8	53.2	58.2	54.5	57.6							
		σ	8.4	9.9	9.8	8.6	7.9	9.5							

Advanced	152	M	58.7	56.2	58.5	61.3	57.3	61.6				51.0	2.1	26.4	10.9
		σ	7.6	8.8	7.8	6.4	6.8	7.9				21.6	.8	4.7	1.6
Primary	202	M	49.4	47.7	51.3	54.3	51.4	51.2				46.7	2.5	25.7	10.0
		σ	7.3	8.6	7.7	6.8	7.8	8.1				18.6	.8	4.9	1.9
For Fire Controlmen															
Fire Control-Advanced															
Graduating Group	100	M	62.6	62.4	61.7	60.3	67.2	66.2							
		σ	7.1	8.6	8.7	7.1	5.4	7.6							
Group I	203	M	59.0	59.2	58.1	59.5	66.0	63.1							
		σ	8.2	10.8	8.7	7.6	6.5	9.5							
Group II	114	M	60.1	60.0	57.2	59.7	65.3	63.0							
		σ	8.8	11.4	8.4	7.4	7.3	9.6							
Group III	72	M	60.5	60.0	59.6	59.4	64.0	63.6							
		σ	8.0	9.4	8.5	6.9	7.3	8.0							
For Water Tenders															
Oil Burning															
Group I	173	M	49.2	48.2	51.9	54.6	51.8		48.6	51.0		42.7	3.1	25.4	10.0
		σ	11.3	11.5	8.7	8.3	8.1		9.1	9.2		22.3	1.0	5.2	1.7
Group II	169	M	48.7	47.2	62.6	54.4	49.7		50.0	53.2		47.0	3.0	26.2	9.8
		σ	8.6	9.3	10.2	10.3	9.3		9.0	9.3		24.8	1.0	5.5	1.9
For Sonarmen															
West Coast Sonar	143	M	61.3	56.0	59.3		59.1				67.2	30.3	3.2	23.8	11.4
		σ	7.7	9.9	7.8		8.7				9.0	7.1	.6	4.7	1.4
Fleet Sonar	111	M	60.7	60.0	59.6		62.7		74.5	63.7		31.4	3.4	23.4	11.4
		σ	7.9	11.2	8.5		9.1		8.5	9.7		10.2	.7	3.8	1.2

[1] Pay grades were given the following numerical values:

Chief Petty Officer—1
First Class Petty Officer—2
Second Class Petty Officer—3
Third Class Petty Officer—4
Seaman or Fireman First Class—5

school mentioned above, the written test grades and the laboratory grades were considered to be equally important. The averages of both types of grades were, therefore, given unit weight in computing the final average. But when the quiz grade averages were correlated with final grades a coefficient of .98 was obtained, while a similar correlation between laboratory grade averages and final grades yielded a coefficient of .55. It is highly improbable that this school was unique in making such errors in procedure.

Considering the characteristics of the criterion data which have just been described, it is not surprising that validity coefficients for prediction variables are often low. Neither should it be surprising that tests, like the General Classification Test and the Reading Test, which are largely measures of verbal ability, often show almost equal relationship to grades in mechanical schools with tests purporting to measure mechanical aptitude. Estimation of the true relationships of predictor variables to some measure of performance awaits research in which the criterion measures receive their due share of attention.

METHODS OF ANALYSIS. The analysis was principally correlational. For continuously distributed variables, such as test scores, product-moment correlation coefficients were computed. Biserial coefficients were used to analyze the relationship between the criterion and attendance or non-attendance in related elementary service schools. The Wherry-Doolittle procedure [1] was employed to select the combination of three tests which would yield the highest multiple prediction.

Results

STATISTICAL DESCRIPTION OF POPULATION. Table 1-XIII presents the means and standard deviations of test scores and other selected data from the Enlisted Personnel Qualifications Cards. Data concerning civilian occupation and prior naval school training are reserved for later presentation.

For most of the school groups, the test score means show a substantial increase and the standard deviations a definite shrinkage from the corresponding statistics for Navy recruits. The groups from NTSch (Oil Burning) are an exception to this. The means and standard deviations from these groups are just about what one would expect from samples of recruits. Likewise, the primary group from NTSch (Gunner's Mates and Electric Hydraulics) shows no significant differences with respect to mean test scores, although there is a definite shrinkage in standard deviations. Differences from

[1] Stead, W. H., Shartle, C. L., et al. *Occupational Counseling Techniques*, pp. 245-252. American Book Co., New York, 1940.

the recruit population with respect to test score performance approach a maximum in the groups from NTSch (Electrical Interior Communications).

The particular system of dividing trainees at NTSch (Gunner's Mates and Electric Hydraulics) into primary and advanced groups accounts for the differences between these two groups shown in the table. Following the earlier study at this school, which was begun in the late summer of 1944, the school's program was reorganized to provide these two different curriculum levels. The recommendations for selection of trainees resulting from the earlier study were followed, as closely as possible, in assigning men to the advanced curriculum. The residue from this selection process, together with men thought unqualified for the advanced work because of deficiencies in background, were placed in the primary course.

Wherever the differences in test score means and standard deviations are substantial, they indicate the operation of some selection process, although distributions for certain tests might be expected to show restriction in range and higher means where men took tests at stages of Navy experience later than recruit training. The extent to which these differences represent selection from the populations available for assignment to the particular schools is not known, however. The test score distributions for these populations, had they been known, might have shown an equal degree of selection as compared to a group of recruits. For example, the men assigned to NTSch (Electrician's Mates) and NTSch (Gyro Compass) appear to be rather highly selected. Their test score means on the General Classification Test, Arithmetical Reasoning Test, and Mechanical Aptitude Test are all well above those for Navy recruits; and the standard deviations are consistently smaller. On both parts of the Mechanical Knowledge Test the differences are considerable. Yet differences of like magnitude could have obtained for all men who were eligible for assignment to advanced electrician's mate schools. It is possible that the eligible electrician's mate group was that highly selected. While it seems unlikely that the electrician's mate sample was a random sample from the available population, the hypothesis must be considered tenable since statistics are not available to disprove it.

The data on "months of active duty" show marked differences from school to school. Length of service was longest (from three and one-half to four years on the average) for the samples from NTSch (Gunner's Mates and Electric Hydraulics) and NTSch (Oil Burning). It was shortest for the factory engine schools (less than a year and a half on the average for NTSch [Diesel]—B). Many of the trainees in these factory schools had only recently graduated from

TABLE 2-XIII. Correlation coefficients of school success with test scores and other selected personal data for samples from twelve advanced Enlisted Schools. The criterion of school success was final school grades, or a linear scale transformation of ranks derived from final school grades.

School	N	Test Score									Other Personal Data			
		GCT	ARI	MAT	MKM	MKE	READ	SPELL	CLER	RADIO	Mo. of Active Duty	Pay Grade[1]	Age in Years	Years of Civilian Education
For Motor Machinist's Mates														
Diesel A	148	.52	.49	.54	.45	.31					.00	.29	.11	.18
Diesel B	107	.25	.32	.37	.27	.37					.13	.16	.15	.21
Diesel C	40	.37	.32	.58	.34	.17					.17	.08	.22	.34
Packard Marine Engine	106	.24	.17	.12	.26	.32		.11	.11	.02			.05	.20
For Electrician's Mates														
Electric Interior Communications														
Group I	119	.28	.35	.28	.09	.07	.24				.23	.05	.12	.14
Group II	109	.26	.38	.23	.28	.22					.06	.11	.12	.10
Electrician's Mates	205	.48	.39	.46	.28	.43					.01	.30	.02	.36
Gyro Compass	59	.43	.46	.51	.44	.37					.09	.35	.08	.12
For Gunner's Mates														
Gunner's Mates and Electric Hydraulics														
1st Study	154	.55	.46	.51	.44	.48	.63							
Advanced	152	.27	.31	.16	.25	.25	.27				.15	.21	.01	.17
Primary	202	.05	.30	.17	.32	.22	.31				.08	.14	.05	.26
For Fire Controlmen														
Fire Control Advanced Graduating Group	100	.32	.32	.26	.36	.54	.38							
Group I	203	.54	.59	.40	.43	.54	.53							
Group II	114	.53	.55	.52	.43	.56	.51							
Group III	72	.45	.59	.31	.39	.48	.53							
For Water Tenders														
Oil Burning														
Group I	173	.63	.66	.53	.50	.45		.38	.39		.37	.54	.19	.33
Group II	169	.45	.42	.40	.47	.32		.32	.23		.50	.65	.15	.19
For Sonarmen														
West Coast Sound	143	.21	.37	.23		.02				.36	.26	.19	.08	.26
Fleet Sonar	111	.19	.38	.34		.03		.16	.22		.08	.21	.08	.23

[1] Pay grades were given the following numerical values:

Chief Petty Officer—1
First Class Petty Officer—2
Second Class Petty Officer—3
Third Class Petty Officer—4
Seaman or Fireman First Class—5

basic engineering schools and had been in the Navy as few as six to eight months. But in all cases the standard deviations of the length of naval service data are large. All of the samples included a wide range of length and type of Navy experience.

Data on age and years of education are also presented in Table 1-XIII. It is apparent that there were also systematic differences between schools with respect to these factors.

RELATIONSHIP OF PREDICTOR VARIABLES TO SCHOOL SUCCESS. Table 2-XIII contains the zero-order validity coefficients of the same predictor variables for which means and standard deviations were presented in the previous table. Since test scores and other factors correlated with test scores were used in selecting men for assignment to certain schools, it is probable that the corresponding validity coefficients are underestimates of the population values. No corrections have been applied to these data, as was done in the elementary school studies, because the necessary knowledge of the population distributions was lacking. Additional data on the relationship of age to school success and on the relationship of prior Navy training and civilian occupational experience to school success are shown in Tables 3-XIII, 4-XIII, and 5-XIII respectively.

APTITUDE TEST SCORES. Among the motor machinist's mate schools, the correlation coefficients are highest for samples from Diesel Schools A and C and lowest for the Packard Marine Engine School and Diesel School B. Reference to Table 1-XIII indicates that the trainees for the latter two schools seemed also to be more highly selected. But the differences in restriction of test score distributions do not appear to be sufficient to account for all of the differences in correlation coefficients. Considering all schools together, the Mechanical Aptitude Test appears to be the best single predictor for the motor machinist's mate schools. It fails to correlate well with success in the Packard Marine Engine School, but so do all of the other tests. The Wherry-Doolittle analysis selected the Mechanical Aptitude Test as one of the most predictive combinations of three tests in each instance, even in the Packard Marine Engine School, where its zero-order correlation with the criterion was below that of four other tests.

The finding that this test, which includes items measuring comprehension of simple mechanical relationships and principles and items measuring facility in spatial manipulation and perception of spatial relationships, predicts well for this type of school is certainly in line with expectation. Indeed, one might expect a test of this sort to enjoy a greater relative advantage compared to tests like the General Classification Test. That it did not is probably due to (1) high intercorrelations with other tests of the Basic Test Battery,

TABLE 3-XIII. Criterion score means and standard deviations for four age groups in samples of trainees from ten advanced Enlisted Schools. Criterion scores are linear scale transformations of ranks derived from final school grades. Data in italics are for categories where N is less than 30. No statistics are given for categories in which fewer than 10 cases were available.

Schools	Age in Years: 17-20 M	17-20 σ	21-24 M	21-24 σ	25-29 M	25-29 σ	30 and older M	30 and older σ
For Motor Machinist's Mates								
Diesel A	44.5	17.8	49.5	19.0	53.8	19.3	51.2	18.0
Diesel B	46.9	16.0	50.4	20.3	49.5	19.9	55.3	19.2
Diesel C	*57.1*	*15.6*	47.7	19.4				
Packard Marine Engine	48.5	17.9	*53.0*	*19.6*	53.8	17.0		
For Electrician's Mates								
Electric Interior Communications								
Group I	*52.3*	*18.5*	46.5	17.5	52.0	17.2	53.1	17.1
Group II	*52.5*	*18.9*	50.5	17.8	52.3	19.0	*42.8*	*12.9*
Electrician's Mates	*47.7*	*18.1*	47.3	18.5	53.4	21.1	47.9	20.7
Gyro Compass	*45.7*	*19.2*	49.6	18.5	53.0	19.6	*53.0*	*9.7*
For Gunner's Mates								
Gunner's Mates and Electric Hydraulics								
Primary	*44.3*	*19.2*	47.3	18.7	53.2	18.5	50.8	21.2
Advanced	*58.0*	*16.5*	47.0	22.5	50.5	20.5	42.1	20.5
For Water Tenders								
Oil Burning	45.3	18.6	46.9	17.2	58.0	19.8	51.4	20.6
For Sonarmen								
West Coast Sound	50.7	17.0	51.1	19.6	48.8	21.5	*46.4*	*20.4*
Fleet Sonar	*49.1*	*24.0*	47.3	17.7	*57.0*	*17.8*	*49.3*	*10.1*

TABLE 4-XIII. Relationship of prior Navy training to success in advanced service schools. (1) Biserial correlation coefficients between criterion and graduation on non-graduation from an elementary Naval Training School. (2) Product-moment correlation coefficients between criterion and grades in prior training.

Schools	Diesel A	Diesel B	Diesel C	Packard Marine Engine	EIC Group I	EIC Group II	Electrician's Mates	Gyro Compass	Oil Burning Group I	Oil Burning Group II	West Coast Sound	Fleet Sonar
Graduates vs. Non Graduates of Elementary Schools (1)	.16	.05	.34		.35	.14	.32	.25	.07	.09	.22	.08
Elementary School Grades (2)	.56	.09	.53	.44	.65	.55	.37	.59	.46	.53	.54	.26

TABLE 5-XIII. Criterion score means and standard deviations of samples from ten advanced Enlisted Schools, subdivided according to civilian occupation groups. Criterion scores are linear scale transformations of ranks derived from final school grades. Data in italics are for categories where N is less than 30. No statistics are given for categories in which less than 10 cases were available.

Schools	Professional and Managerial		Clerical and Sales		Service		Agriculture, Fishing, Forestry		Manual, Skilled		Manual, Semi-skilled		Manual, Un-skilled		None	
	M	σ	M	σ	M	σ	M	σ	M	σ	M	σ	M	σ	M	σ
For Motor Machinist's Mates																
Diesel A	*56.5*	*12.3*	*49.3*	*20.0*			*60.0*	*16.2*	54.0	18.3	49.8	20.0			44.6	16.1
Diesel B							*58.3*	*17.2*	52.0	18.8	44.7	20.2			47.8	19.1
Diesel C									*42.2*	*18.7*	*52.9*	*13.7*			*53.9*	*16.9*
Packard Marine Engine							*52.8*	*17.7*	*53.9*	*17.9*	*47.2*	*15.1*			*38.7*	*14.4*
For Electrician's Mates																
Electric Interior Communications[1]																
Group I	*59.8*	*12.6*	*65.6*	*14.6*					57.4	13.5	*53.9*	*13.4*			57.4	14.2
Group II			*51.0*	*20.3*					52.8	18.2	*52.0*	*20.9*			46.5	17.0
Electrician's Mates			*44.1*	*23.4*			*48.7*	*17.8*	52.1	19.5	46.3	19.6	*57.7*	*13.2*	49.8	19.3
Gyro Compass									*52.9*	*20.9*	*47.1*	*12.4*	*52.1*	*18.1*	50.9	21.3
For Gunner's Mates																
Gunner's Mates and Electric Hydraulics Primary			*53.0*	*18.1*	*45.9*	*16.5*	*53.6*	*20.0*	45.2	19.0	44.5	20.6	*51.8*	*20.4*	51.8	18.1
Advanced			*45.3*	*18.4*			*55.1*	*17.3*	*48.6*	*18.9*	45.6	17.0			46.2	25.5
For Water Tenders																
Oil Burning			*49.2*	*18.4*			46.4	21.6	50.0	19.0	49.2	17.9	*44.4*	*17.4*	53.6	18.4
For Sonarmen																
West Coast Sound	*57.4*	*15.8*	*46.5*	*17.7*					*52.7*	*14.6*	*48.2*	*21.5*			46.7	15.2
Fleet Sonar			*50.1*	*17.8*					*50.8*	*10.2*	*48.7*	*19.7*			48.4	21.0

[1] For this group, no occupational experience data were available for failing students (approximately 20 per cent). Since the failures were included in assigning criterion ranks, the means for occupational categories are based on only the upper 80 per cent of the criterion ranks, and are, therefore, higher than the corresponding means for other groups.

(2) weaknesses in the criterion, including failure to evaluate accurately and give proper weight to laboratory performance, and (3) deficiencies of the test itself as a predictor of ability to perform work of a mechanical nature at the advanced school level.

For two of the three electrician's mate schools, NTSch (Electrical Interior Communications) and NTSch (Gyro Compass), the best correlation with success was obtained for the Arithmetical Reasoning Test. The curricula of both of these schools are heavily weighted with mathematics. At NTSch (Electrician's Mates), with a somewhat less mathematical curriculum but with a considerable body of work in laboratory repair and maintenance, the General Classification Test and Mechanical Aptitude Test appear to be best. The Mechanical Aptitude Test possesses some predictive capacity for all of these electrical schools. A Wherry-Doolittle analysis to select the best combination of three tests includes it in each instance.

The group labeled First Study from NTSch (Gunner's Mates and Electric Hydraulics) is the only one from that school yielding data that can be compared with data from the other schools in the table. Data for the primary and advanced groups from this school were collected after the recommendations from the first study had been put into effect, and are interesting principally as a demonstration of what happens to validity coefficients when test score data are obtained from groups that have been systematically selected on the basis of those same test scores. In the first study the Reading Test showed an unusually high correlation with the criterion. Unfortunately, not enough scores were available on this test from most of the schools to provide much comparison data. The recommended combination of tests, from this First Study, was the Reading Test and the Mechanical Aptitude Test.

The advanced fire control sample was subdivided into four groups on the basis of the stage of training at which the men were tested. Relationships between the predictor variables and the criterion were quite stable for the last three groups. The first group varied considerably from the other three in composition. It included a considerably greater proportion of men assigned from elementary service schools and fewer from sea duty, and it was generally a more systematically selected group. The two tests that were recommended for selection of trainees for this school were the Arithmetical Reasoning Test and the Mechanical Knowledge Test (Electrical Score).

The correlation coefficients for the Oil Burning School vary considerably between the two classes. To some extent these variations correspond to the magnitude of the standard deviations of the corresponding tests. But not enough is known concerning the composition and characteristics of the two groups to account completely for the differences. Taking both groups together, the General Classi-

fication Test and the Arithmetical Reasoning Test are probably best. Wherry-Doolittle analysis selects these two tests as part of the best combination of three for both groups. In Group II the Mechanical Knowledge Test (Electrical Score) yields a slightly higher relationship than either of these two, but it is relatively less efficient for Group I.

In the two sonar training schools the only tests that were found to be promising for prediction were the Arithmetical Reasoning Test and the Radio Code Test—Speed of Response. The Mechanical Aptitude Test showed a fair degree of correlation with the criterion for the Fleet Sonar School group, but a considerably lower coefficient was obtained for the West Coast Sound School sample. The correlation obtained for the Arithmetical Reasoning Test may be due, in part, to the emphasis on plotting and interpretation of sonar data for navigation and fire control problems in the advanced sonar curriculum. Unfortunately no data for the Radio Code Test were available from the Fleet Sonar School, so the correlation obtained for that test in the West Coast Sound School cannot be checked. It is unfortunate that Sonar Pitch Memory Test (Chapter VIII) scores were not available for these trainees. Certainly, if special selection tests for special types of training have a function anywhere in the Navy's selection program, it is for such psychologically unique types of tasks as those encountered in sonar operation.

Navy Experience. The correlations of "months of active duty" and "pay grade" with the criterion are low and of no significance for prediction, except for one school, NTSch (Oil Burning). The reasons for this unique finding are not known, but at least two possible explanations may be offered. The first is that this was the one school of the group studied in which the materials taught, methods of instruction, and methods of measuring achievement were of such a nature as to give a real advantage to the men with long experience and higher rates. The second possibility is that the grades assigned in this school were contaminated by the influence of prestige factors associated with experience and high rating. The thorough investigation of the methods of instruction, achievement testing, assignment of grades, etc., that would be needed to reveal the correct explanation, was not undertaken in this study. Such investigations will need to be routine for future studies if these variables are to be evaluated. It may be noted in passing that even in the Oil Burning School the intercorrelations of "pay grade" and "months of active duty" with Basic Test Battery scores were low—of the order of .10 to .20.

Age. The correlation coefficients between age and school success are low and variable. In a few cases low negative coefficients were found. But it is possible that the product-moment correlations

underestimated the real relationship, since the regression of age on school success often tends to be curvilinear. For this reason a further analysis was made. The samples were divided into four age groups, and separate distributions of criterion scores for each age group were constructed. The means and standard deviations of these distributions may be found in Table 3-XIII. For most schools the data in this table bear out the hypothesis of a curvilinear relationship. With a few exceptions, the mean scores show an increase from the youngest group (17-20 years) through the third group (25-29 years), and no increase or a decline for the fourth and oldest age group (30 and above). But the distributions for the separate age groups overlap considerably (note the large standard deviations). Inspection of the distributions therefore indicated that the relationships were very low, and the *etas* were not computed. So far as those data can be considered indicative, age does not appear to be of very great predictive value for selecting men for advanced training.

AMOUNT OF CIVILIAN EDUCATION. With the exception of one group, years of education correlated positively with the criterion. In some schools the coefficients appear high enough to be of possible value in selection. But the intercorrelations between Basic Test Battery scores and years of education are, in general, so high that the inclusion of the latter variable adds little to prediction. In only two instances did the Wherry-Doolittle analysis select this variable as one of the most predictive combination of three. The two samples in which this occurred were the Packard Marine Engine School and the primary group of NTSch (Gunner's Mates and Electric Hydraulics). In both cases correlation coefficients between the Basic Test Battery scores and the criterion are found to be generally low. In samples where reasonably good prediction could be obtained from test score data, the variable of years of education added nothing. It might be possible to find relationships between type of education (number of shop courses, or mathematics courses, etc.) and success in certain advanced enlisted schools which would be valuable for prediction purposes. These studies do not provide data with which to test this hypothesis.

PREVIOUS NAVAL TRAINING. Studies of prediction of scholastic success in civilian schools have demonstrated that one of the best predictors of an individual's likelihood of success in school is knowledge of how well he has previously succeeded in school. Table 4-XIII presents data concerning the relationship of previous naval school training to success in advanced schools. For all schools except NTSch (Electrical Interior Communications), the prior Navy school training was at the elementary level—class "A" school, or basic engineering. Biserial correlation coefficients were computed to determine the

relationship between graduation from an elementary school and success in advanced school. These are the coefficients in the first row of the table. For the most part the elementary school graduates seem to have enjoyed only a slight advantage as a result of their previous school training. In at least half of the cases it was negligible.

Correlation coefficients between elementary school grades and success in advanced schools for the elementary school graduates are found in the second row of the table. With one exception, a reasonably good degree of relationship is indicated. The obtained correlations are probably attenuated by (1) variable time lapse since graduation from elementary schools, (2) lack of close relationship between elementary and advanced school curricula, especially for men who were not recent elementary school graduates, and (3) unreliability of grading systems. The practical usefulness of this relationship is limited by the small proportion of eligible advanced school candidates who have attended elementary schools. The relationship might become important, however, if training programs should be developed in which most or all of the training is done in schools, rather than on the job. In such a program, the selection of men for advanced training might well weight heavily the factor of performance in elementary training.

The Electrical Interior Communications School represented a unique case of this type of selection. All trainees for that school were selected from the men who had been first assigned to NTSch (Electrician's Mates), during the first two weeks of which the only subject studied was mathematics, primarily arithmetic and elementary algebra. At the end of this two week period the group to be transferred to Electrical Interior Communications School was selected and segregated for an additional six weeks in a speeded-up preparatory course. During this time basic d-c theory, in addition to more mathematics, constituted the course of study. Accordingly, the following Electrician's Mates School grades were obtained for the EIC School trainees: (1) average grade in mathematics at end of first two weeks, (2) average grade in mathematics at end of 8th week, (3) average grade in d-c theory at end of 8th week, (4) 8 weeks average including both mathematics and d-c theory.

Correlation coefficients between these grades in Electrician's Mates School and final grades in EIC School are as follows:

	Group I	Group II
Mathematics—2 weeks average	.23	.40
Mathematics—8 weeks average	.41	.61
D-c theory—8 weeks average	.66	.79
Mathematics plus d-c theory—8 weeks average	.62	.77

Since the selection occurred at the end of the first two weeks, the two weeks' mathematics grade was the principal selection variable. This procedure was dictated in part by administrative considerations and also by the belief that the most necessary qualification for mastery of the EIC School curriculum was aptitude for mathematics. As a matter of fact, the data show a higher correlation for the grades in d-c theory than for mathematics. This difference between correlations should be interpreted with care, however, because the range of grades in mathematics was restricted as a result of the use of mathematics grades as the selection variable. In any case, use of a longer base than is provided by a two weeks grade is amply indicated.

CIVILIAN OCCUPATIONAL EXPERIENCE. The civilian occupation data are presented in Table 5-XIII. Distributions were constructed for each of the main occupational categories, as indicated by the first digit of the occupational codes. Statistics that are relatively unstable because of the small number of cases (N less than 30) are indicated by italics. No statistics are given for those categories in which there were fewer than ten cases.

Inspection of the table indicates very little possibility of prediction from these data. For example, there are such anomalous findings as the following: (1) At NTSch (Gunner's Mates and Electric Hydraulics) one would expect men with experience in mechanical occupations, especially skilled workers, to have an advantage. Yet the data indicate that men with codes indicating such occupational backgrounds made lower grades on the average than men coded as having experience in agricultural occupations or in clerical and sales occupations, or with no civilian occupational experience. (2) In the diesel schools there was only a very low relationship between coded level of skill in manual occupations and school success. In one case the semi-skilled group averaged ten points higher than the skilled group. Both groups tended to be below the agricultural worker group.

The unsatisfactory character of these data has been discussed in an earlier section. One additional point should be made. The samples from the separate occupational categories were systematically biased by the use of occupational experience as a selection factor. Men with civilian occupational experience that was considered to be related to some type of Navy job were assigned to training under less demanding standards of test score performance than were applied to men with experience in "unrelated" occupations or with no occupational experience. In addition the most promising men were taken from the groups with "related" occupational experience for Navy jobs that were considered to be most demanding of knowledge and skill, or to have highest priority. Navy jobs with lower priority

may have received the residue from the "related" occupational experience groups. As a result it is likely that, at least in some schools, a generally higher class of men were assigned from the groups with no civilian occupational experience, or with experience considered unrelated to the training, than from the groups with "related" experience.

For these reasons no valid estimates of the relationships of civilian occupational experience to Navy jobs, or success in training for Navy jobs, can be secured from the present data. The appropriate procedure for obtaining this information would insure either (1) random sampling within each occupational group whose relationship it was required to evaluate, or (2) collection of data that would enable comparisons to be made between equivalent groups, either by suitable fractionation of data, or by applying suitable corrections to criterion data. In general, samples would need to be much larger and more carefully designed than was the case in these studies.

Summary

Studies were conducted in a number of advanced service schools for enlisted personnel to determine the relationship of available qualifications data to success in advanced training as measured by the grades assigned by the schools. Although the studies were limited by various contingencies arising from wartime conditions, certain tentative conclusions are possible.

The results show that test score data can be used to predict the school grade criteria employed in these studies with, on the whole, reasonably good success. For schools whose curricula require mastery of a considerable amount of mathematics or facility with computational procedures, some test of mathematical aptitude, such as the Arithmetical Reasoning Test, has significant predictive value. For schools concerned with training men for jobs of a mechanical nature, or whose curricula include a considerable amount of laboratory work, the Mechanical Aptitude Test was useful in prediction. At the advanced school level the Mechanical Knowledge Test (both Mechanical and Electrical Scores) are, in general, relatively less efficient predictors than for elementary schools. For most of the schools some measure of verbal facility (the General Classification Test or the Reading Test) showed substantial correlation with the criterion. Although ability to master written materials is no doubt important in most school situations, part of this correlation may result from the fact that grades were determined principally from paper and pencil tests in which the verbal factor is heavily weighted.

Grades in previous Navy training correlated well with grades in

advanced schools. Although not applicable to many selection situations, this relationship should make possible good prediction where it can be applied.

"Months of active duty" and "pay grade" correlated highly with the criterion for the groups from NTSch (Oil Burning) but failed to correlate significantly in any other school. No satisfactory information to explain this anomalous result was available from these studies.

Correlation coefficients of school success with age were too low to be of value for prediction.

Although number of years of civilian education showed somewhat higher correlations, multiple validity coefficients were not appreciably increased by the inclusion of this variable, owing to the high intercorrelations between years of education and Basic Test Battery scores.

No determination of the relationship of civilian occupational experience to success in schools could be obtained from the studies reported here. The data were unsatisfactory because of lack of standard recording procedures and because the samples of various occupational groups were biased.

PART IV

THE CONSTRUCTION AND USE OF ACHIEVEMENT MEASURES

CHAPTER XIV

SERVICES PROVIDED TO NAVY TRAINING THROUGH ACHIEVEMENT EXAMINATIONS

THE use of standardized achievement examinations in the Navy has been almost exclusively the result of activities during the calendar years 1944 and 1945. Prior to that time, locally made examinations had been prepared when they seemed to be required. These reflected the curriculum as it was taught in the particular school and the concepts of examining held by the individual instructor. As the work of the Standards and Curriculum Division in standardizing school curricula began to take hold, the desirability of a common measure of the results of instruction in a given type of school became increasingly apparent. Official cognizance of the need was taken during the latter months of 1943 when a request was made by the Quality Control Division of the Training Activity for standardized measures of school attainment. By January 1, 1944 two officers and a yeoman had been assigned to the construction of achievement tests. Maximum strength of the program was reached during the first half of 1945.

As indicated above, the development of the Navy's achievement testing program was probably most closely associated, historically, with the efforts to standardize curricula and instruction in Navy schools. Prior to 1942 the curricula followed by the several schools of a given type varied extremely, although the instruction in each was presumably directed toward certain general objectives established by the Bureau and based upon requirements dictated by experience at sea. To overcome in part the diversity in trainee attainment, the preparation of uniform courses of study was undertaken. These standardized curricula set forth certain minimum essentials for all schools of a given type and represented a significant step forward toward the provision of naval personnel who would have common backgrounds of understanding and skill in certain specialized jobs. This permitted the necessary interchangeability of manpower demanded by the exigencies of sea duty. The standardization of curricula, however, could not in itself assure the attainment of this objective. There was always the possibility, for example, that instructional materials and practices would not conform in the desired manner. Regular visits to naval training schools by representatives of the Bureau of Naval Personnel provided a certain degree of supervision of instruction but did not insure strict

adherence to the prescribed courses of study. Furthermore the standardized curricula alone could not guarantee a standardized product —the graduation of trainees possessing comparable amounts of knowledge, skill, and general competence in Navy ratings. The identification of the differences that existed between trainees after training became a problem of major concern.

The achievement testing program was undertaken, therefore, primarily for the purposes of (1) contributing to the standardization of instruction in Navy schools, and (2) improving the bases for comparing school trained personnel.

To meet the needs implied in these stated purposes (the need for a means of assuring teaching of minimum essentials of the course, the need for reliable bases for marking and assigning grades, and the need for objective comparability of school graduates) the Test and Research Section undertook to provide written and performance examinations for training activities. These tests were planned to sample representative knowledges and skills of a number of courses of instruction.

It is significant that although the formulation of these achievement tests and performance problems was guided by the standardized curricula, attention was constantly focused on shipboard duties and functional knowledge. The requirements of the billets in which school graduates would serve upon completion of training were kept constantly in mind, and every effort was made to relate to shipboard practice both written examinations (through the extensive use of drawings and verbally described situations requiring application of information to practical problems), and performance tests (requiring actual operation of gear and demonstration of skills required in the job).

Types of Activities for Which Examinations Were Provided

With a limited number of personnel available during the period when demands for officers in combat areas were great, it was not possible to develop examinations for all Navy training simultaneously. As a matter of expediency it was agreed that the elementary enlisted training schools should receive first attention. Two criteria were employed in determining which schools would be considered first in the development of achievement tests: (1) the number of trainees regularly assigned (which reflected the billets in which the needs for personnel were greatest) and (2) the overall significance of the type of training offered. In light of these criteria the construction of achievement examinations was first undertaken for the enlisted schools for gunner's mates, electrician's mates, yeomen,

storekeepers, motor machinist's mates (diesel) and signalmen. During the period from January 1944 to July 1945, performance tests and written examinations were prepared in from 2 to 6 comparable forms for the following enlisted schools and courses:

Class P, NTSch (Basic Engineering)
Class A, NTSch (Diesel)
Class A, NTSch (Electrical)
Class A, NTSch (Fire Controlmen)
Class A, NTSch (Gunner's Mates)
Class C-1, NTSch (Gyro Compass)
Class A, NTSch (Quartermaster)
Class P, NTSch (Radar Operator)
Class A, NTSch (Radio)
Class A, NTSch (Signal)
Class A, NTSch (Storekeepers)
Class A, NTSch (Torpedomen)
Class A, NTSch (Yeomen)
Lookout Courses
Telephone Talker Training

Of special note is the program of achievement examinations in its relation to training in the new and significant area of electronics. Multiple forms of examinations were prepared for pre-radio materiel and for elementary electricity and radio materiel schools. This examination program is described in Chapter XVII.

Chronologically the development of achievement examinations for officer schools followed that of the enlisted schools. During late 1944 and early 1945, standardized examinations in navigation, seamanship, ordnance and gunnery, and damage control were prepared and administered in the naval reserve midshipmen's schools (deck course). Officer examinations were also prepared covering subject matter for the pre-radar course of instruction and for tactical radar training.

Examinations for recruit training were projected from the beginning of the achievement testing program. They were prepared and administered in experimental form in January 1945 with subsequent revision for routine use at completion of recruit training. The examinations were comprehensive, consisting of sections of items representing the various areas of the recruit curriculum, namely, naval organization, seamanship, military training, ordnance and gunnery, fire fighting, gas warfare defense, recognition, lookout, telephone talker, and first aid and personal hygiene.

From time to time, assistance in achievement examining was pro-

vided to additional training activities. Two forms each of a Reading Classification Examination, for determining pre-instruction reading abilities, and a Reading Achievement Examination, for measuring the end-of-course attainment, were prepared for and used in the Navy's special training program for illiterates. A number of performance and written examinations were constructed also for use in connection with the important amphibious training program under the direction of the Amphibious Training Command, Pacific Fleet.

The Improvement of Instruction and Learning

The services provided to Navy training through achievement testing were in direct response to the needs which gave rise to establishment of an achievement examination program. They were, first, those leading to improvement of instruction and learning, and second, those having to do with improvement of grading and marking.

One of the original purposes of the administration of standardized achievement examinations was the improvement of instruction through standardization of curricula in all schools of a given type. The examinations, based upon prescribed curricula for the several types of schools, were administered regularly in accordance with directives from the Bureau of Naval Personnel. Following administration of an achievement examination, the answer sheets were scored and the results converted into achievement examination grades for use in combination with regular course marks in compiling the trainees' final course grades. These achievement examination grades were empirically established to permit direct comparability from school to school and from one form of the examination to another. Answer sheets together with distributions of achievement examination grades were forwarded by the school to the Bureau of Naval Personnel, where measures of central tendency and of variability were computed for each graduating class of each school. The distribution of grades and other pertinent statistics were forwarded to the Quality Control Division and the Instructor Training Section of the Training Activity to be used for supervisory purposes (1) in determining the extent to which schools appeared to be accomplishing their instructional purposes, and (2) in recommending revisions of curricula, methods, and policy as seemed desirable. It was thus possible, through administration of the achievement examinations, to obtain objective estimates of the relative effectiveness of the schools' programs, at least to the extent that the outcomes are reflected in trainee attainment on these examinations.

While the use of achievement examination results for supervisory purposes constituted an important service, and while this was one of the basic purposes for which the achievement testing program was inaugurated, the local use of the examinations was perhaps even more significant in the overall improvement of instruction and learning. Early in the program of test development, representatives of the Test and Research Section found that they could most effectively produce appropriate testing materials by working directly with instructors in the schools. This practice not only permitted the test technicians to develop close acquaintance with the curriculum and skills taught and to obtain valuable assistance from staff members of the schools in the construction and try-out of testing situations, but it also enabled them to lend the school, in return, useful services by helping to interpret test results, and by indicating subject matter areas in which the instruction or curriculum appeared to be inadequate and in need of attention. Later in the program, this service became a special responsibility of field representatives of the Instructor Training Section assigned to school commands, although officers of the Test and Research Section continued to provide advice and assistance as required.

Special detailed analyses of the achievement examination results were prepared by the Test and Research Section for both recruit training commands and the reserve midshipmen's schools (deck course). These analyses, together with adequate interpretation and suggestions for their utilization, were reported to each of the activities concerned as guides for the improvement of instruction. Similar analyses, showing the percentages of graduates of a given school responding correctly to each item or problem, were prepared and generalizations formulated to show the topics and problems that were being less adequately presented.

In addition to the assistance provided to schools by item analyses, participation in the development of examination materials served to stimulate and guide instruction. It has been a generally accepted principle in education that the experience of devising examination questions in areas of study aids materially in the understanding and clarification of the concepts involved. Instructors in naval training programs who participated in the construction of achievement examinations were able to better identify possible misunderstandings of trainees and to adapt instruction to the trainees' level of ability. The development and use of performance tests was particularly effective in bringing about the improvement of instruction and learning opportunities. Despite the fact that the skills required of naval personnel are usually of a distinctly practical nature, instruction was sometimes dominated by lectures and verbal descriptions

of apparatus and its operation, or at best by demonstrations which were assumed to be effective in acquainting the trainee with the necessary techniques and understanding. Even when learning by doing was undertaken, its purpose was often defeated by lack of the individual supervision and attention required to avoid wrong learning. The introduction of performance tests immediately focused attention upon familiarity with equipment and on the skills and understanding of procedures necessary for maintenance, operation, and repair.

It was a common experience in the naval training schools that when trainees first were required to demonstrate individually their grasp of procedures or proficiency at handling equipment, the results were notably unsatisfactory. But after the routine administration of performance tests had been established, a marked improvement was evident. Testimony of schools' officers indicated that the improvement stimulated by the testing situation was contributed to by a number of factors, including more carefully planned and executed instruction, increased opportunities for trainees to become familiar with equipment and procedures, and more adequate individual supervision of training. The extent to which this change was due to increased motivation on the part of the trainee or improved instructional practices on the part of the instructors is, of course, difficult to estimate.

The Improvement of Grading and Marking

A second major service provided to Navy training by introduction of achievement examinations has had to do with the improvement of grading and marking in the schools. To the extent that the grades or marks are used in the classification of personnel for future assignment to billets, and to the extent to which grades are employed in personnel research, any improvement over the subjective and unreliable estimates characterizing the judgments of untrained teachers would be of considerable importance.

The problem of improving grades and marks as a basis for personnel classification and research has been primarily one of educating instructors in the techniques for improving testing and marking practices. In the first place, it was necessary to point out the advantages of objective course tests and to train instructors in their construction and use. As a result, subjectively assigned notebook grades and marks based on observation of operation of equipment, the daily and weekly essay examinations, and poorly constructed true-false tests were gradually supplanted in many schools by carefully planned ratings of performance (reduced insofar as possible

to a "can do", "can not do" basis) and to comprehensive achievement examinations capable of completely objective scoring.

An outstanding example of the improvement of grades through the development of objective methods of achievement testing was found in one of the basic engineering schools. Here it had been discovered that the Arithmetical Reasoning Test of the Basic Test Battery predicted most effectively the final grades of trainees. While arithmetical reasoning played a significant part in the basic engineering curriculum, it was somewhat surprising to find it so closely associated with success in such training while scores in the Mechanical Knowledge and Mechanical Aptitude Tests showed little relationship to attainment in this program. When objectively scored performance tests were introduced, the dispersion of the grades on shop work was increased and the relative weights of the parts of the course were adjusted to approximate more nearly those that had been assumed to be operating. Subsequent studies of prediction of grades showed the Arithmetical Reasoning Test of the Basic Test Battery to be reduced somewhat in its prognostic value, while the tests of mechanical ability assumed a place of greater importance as predictors of success (for details of this study see Chapter XV).

The necessary training of instructors in the use of objective tests and test results was accomplished largely through the efforts of representatives of the Test and Research Section carrying out temporary duty in the various training commands and also through the development and distribution of a specially prepared bulletin, *Constructing and Using Achievement Tests.* These representatives on duty in the schools were continually alert to the problems of measurement which confronted the instructors and assumed responsibility for the development of understanding of elementary concepts related to tests and handling of test results. Considerable time was spent individually with instructors in assisting them in developing tests, assigning reliable grades, and in reviewing and evaluating their efforts.

The manual, *Constructing and Using Achievement Tests,* was of special value in this respect and was widely used throughout naval training activities. This manual assumed no knowledge or special interest in problems of measurement on the part of the reader. It presented the basic techniques of test construction in simple nontechnical language and in concise form. Four major sections dealt with performance tests, written tests, test administration, and the scoring and interpretation of tests. Emphasis was placed primarily upon the development of tests for local use, particularly performance tests. Examples of satisfactory performance test situations, together with complete directions for setting up equipment, administration,

and scoring were provided. In a section on written tests, the more common types of achievement examination items were presented and fundamental rules and principles were illustrated by example. The section on scoring and interpretation of test scores was particularly useful in the improvement of grading and marking in that it provided a simple linear conversion procedure for translating test scores into grades (Appendix E-2).

A special effort was made to improve marking in specific schools through an examination of grade distributions. Analyses of distributions of schools' grades were made by the Test and Research Section and reports based thereon sent to the schools. Records of the grades assigned to graduates were kept for each school, and distributions of these grades were prepared, together with basic statistics. Comments on the apparent satisfactoriness of the distributions and suggestions for modification of grading standards were forwarded periodically to each school.

Plans for the Future

The achievement testing program during the war was necessarily restricted to those schools or training activities where it was felt the services would be of greatest use. As the program expanded it became possible to serve a growing number of activities and to provide more effective measuring devices. It is expected that, with achievement testing assuming a recognized place in the Research Activity of the Bureau of Naval Personnel, the services in this field will be extended to all types of schools and training regardless of the numbers of trainees involved or the relative significance of the subject matter. It may be expected that a developing program of achievement testing will do much to improve the products of instruction and to assure a well-trained naval personnel.

CHAPTER XV

ACHIEVEMENT EXAMINATIONS FOR ELEMENTARY ENLISTED SCHOOLS

THE major objectives of the achievement testing program have been outlined in Chapter XIV. As stated in that chapter, the majority of examinations built by the Test and Research Section were designed for use in elementary enlisted schools. It was the responsibility of these schools not only to indoctrinate great numbers of men in the ways of the Navy but also to give them the fundamental knowledge and skills expected of them in the many specialized billets they would be required to fill. Ships were manned by great numbers of these new trainees, and the program was variously praised and criticized. There was obviously a need to evaluate how well the program was serving its purpose. The specific questions being asked, which achievement examinations were in part designed to answer, are as follows:

(1) *Are the schools giving adequate training on all aspects of their curricula?* Assuming that the curricula were planned to meet the requirements of the fleet, are the trainees actually learning both the principles and the skills that could reasonably be expected of enlisted school graduates?

(2) *Are enlisted trainees being properly evaluated* with regard to their probable proficiency in the billets to which they will be assigned? Could the prediction of which men will be most successful in each type of training be improved?

(3) *Are parallel schools of the same type giving uniform training?* This is an important problem peculiar to the basic enlisted schools. The enormous numbers of trainees had to be accommodated in many classes in several widely separated schools. The officers and instructors of these schools often had little or no teaching experience to guide them in translating the school curricula into classroom procedures. It was not surprising to find wide variations of emphasis in training, but evidence was needed to show whether there were serious differences in achievement of men trained in supposedly parallel schools.

(4) *Can training be improved by a standardized achievement testing program?* It seemed likely that the testing program would aid training if it were developed with the cooperation of the schools for the purpose of helping the schools evaluate their strong and weak points. One of the most important functions of the testing program proved to be the stimulation of instructor training.

Benefits of the Achievement Examining Program

In order to answer these fundamental questions it was necessary to develop new tests to measure the effectiveness of training. The written tests typically used in the schools measured verbal knowledge, but in many cases they did not effectively measure familiarity with equipment or the ability to use it. Final grades were not entirely satisfactory for evaluating the school training, since grading practices varied from one school to another, and practical ability was not adequately reflected in final grades. The new achievement examinations were developed with emphasis on practical ability and mechanical comprehension. These tests, given under uniform conditions in all parallel schools, provided more dependable standards for measuring and comparing the achievement of men and of schools. But even more gratifying was the observation of beneficial effects on the training program in the following ways:

(1) *Improvement of shop training methods* with more emphasis on training men to *do* required jobs and to *practice* with equipment.

(2) *Standardization of the curricula* by requiring men from all parallel schools to meet similar standards of performance and by revealing deficiencies of training at any one school.

(3) *Improvement of motivation* of students and instructors by setting realistic goals to show trainees their progress in learning the duties of their desired rate. Friendly rivalry for skill became common in several schools.

(4) *Training instructors* to make better weekly examinations, to make more effective use of available equipment by careful scheduling, and to devise interesting training techniques. Good techniques developed in one school could be reported to other schools by the traveling technicians, with a resultant healthy merger of the best ideas.

(5) *Improvement of grades* by giving more weight to practical ability and by using objective test scores to minimize the differences in grading methods between instructors. This gave a better criterion measure for evaluating the predictions made from the Basic Test Battery, and the grades became more useful to classification officers in assigning men for further instruction or for duty billets.

Types of Achievement Measures Developed

PAPER-AND-PENCIL EXAMINATIONS. Printed multiple-choice examinations can be used to measure many results of Navy training, such as mastery of terminology, skill in interpretation of signals, knowledge of correct procedure in sending radio messages or in handling

ordnance, and understanding of the functions of engine parts. Printed booklets of such test items have been prepared for all elementary schools with substantial enrollments. Many items refer to diagrams or to pictures of equipment used in training, with questions about the function of certain parts (Figure 1-xv). Wherever possible, items measure knowledge of "what to do" or how to interpret observations.

The items can be answered on separate machine-scored answer sheets. This greatly speeds the scoring when large groups are tested and permits repeated use of the test booklets. The main advantages of the printed tests are:

(1) ease of administering and scoring
(2) possibility of extensive sampling of topics by using from 75 to 200 items in a one- or two-hour test.

PERFORMANCE TESTS. Since ability to operate and maintain equipment is an ultimate criterion of success in most Navy ratings, performance tests of sample jobs were developed for several programs where the need was most apparent. For each job the trainee is confronted with a set of equipment, in a specified condition, and is told to do something definite with the equipment. Examples for various types of training follow.

Gunner's Mate:

"Remove and replace the extractors."
"Replace the feed pawl and adjust the feed control rod."

Electrical:

"Test these Synchro Units."
"Wire the switches and lamp as a darken-ship circuit."

Basic Engineering:

"Start and check this Gray Marine engine."
"Start this centrifugal pump and cut in the pressure governor."

Torpedoman:

"Calibrate the depth spring for a depth of 10 feet."
"Disassemble the K-gun breech mechanism for cleaning, and reassemble it."

Radioman:

"Tune this TDE transmitter for 6,538 kilocycles."
"Determine the true bearing of the station you will find transmitting at 1,242 kilocycles." (With DAE direction finder.)

Amphibious Training:

> "Make the daily boat-and-engine check. If you find anything wrong tell me what should be done to correct it."
> "Steer a compass course of 190." (Using pivoted boat model with compass.)

While the trainee does the job, a proctor watches him, marking him right or wrong on each step or item on a check-list. Items may be weighted according to their importance, and when it seems desirable, bonuses may be given for speed of performance.

To economize on time, the performance tests are usually administered in batteries of from five to fifteen test jobs, sampling various aspects of training. The trainees move from one set of equipment to another until they finish all the tests of the battery. Multiple sets of equipment are utilized so that more men can be tested simultaneously. Often the performance tests are coordinated with other tests so that all the men are kept busy during a period of several hours. Figure 2-xv shows men working on three 20mm. gunnery performance tests. Examples of directions for administering and scoring such tests are included in Appendix D-1.

Performance tests are the most convincing measure of what a man can actually *do* and of his practical understanding of what he has been taught. The disadvantages are that the tests take much more equipment and more supervision than do paper-pencil tests. Because they are more time-consuming, one is limited to taking fewer samples of a man's achievement.

Identification Tests. In some billets it is important for the men to recognize parts of equipment in order to adjust, repair, or replace them. Identification tests are useful for measuring familiarity with actual equipment rather than with diagrams of it. One procedure has been to arrange tables in some continuous pattern and to lay out, at intervals, the disassembled parts of such mechanisms as guns, pumps, or torpedoes. With each part is a card containing (1) a list of four or five *names,* one of which is the correct name of the object, the others being plausible but incorrect, and (2) a list of four or five *functions* (or methods of adjustment or repair), only one of which is correct (Figure 3-xv). The trainee examines the part and indicates on his answer sheet which of the listed names and functions are correct. He may find it helpful to handle the piece and see it from different angles. At a signal, each trainee then moves to the next part, repeating the procedure. One instructor can test as many men as there are parts in the test (usually from 25 to 50) in a period of from 40 to 80 minutes, allowing time for directions. And if several tests are combined, or duplicate sets of parts used, several

hundred men can be tested at the same time with relatively few proctors. Another procedure, for large assembled units, is to attach numbered test cards to particular parts of the equipment. The trainees move from one part to the next, answering the test items. The identification method has also been adapted to testing knowledge of the purpose of knots, tools, and signals. The method is rapid, interesting to trainees, and a good measure of their familiarity with actual equipment.

PRODUCT-RATING GAGES. In hand-tool or machine-tool shop work, the trainee is required to make metal products to certain specifications of size, squareness, and symmetry. The accuracy of the finished product is taken as a measure of what he has learned about laying out work and using tools correctly. To improve the reliability of grading such products, several special gages have been developed. Figure 4-xv shows two of the gages used in basic engineering schools. The taper gage is used for measuring the diameter of a lightening hole, its symmetry, and the squareness of the sides. The score to be assigned for the diameter is inscribed on the sections of the gage, penalizing the student one point for each .05 inch error in diameter. The movable-pointer gage measures the distance between two shoulders on a machine-turned sample. Scores to be assigned are inscribed on the scale of the pointer. Instructors have found that they can score products more rapidly and accurately with the aid of such gages.

Procedures Followed in Developing Achievement Tests

PAPER-AND-PENCIL EXAMINATIONS. The development of paper-and-pencil examinations followed the procedures usual in civilian educational programs. A test specialist studied the school curriculum, the manuals of instruction, the tests currently in use, and billet analyses revealing the objectives of the training. He conferred with petty officers from the schools to get a better idea of the importance of various topics, and usually had a well-informed petty officer assigned to help in constructing test items, supplying plausible foils for the correct answers, and editing for appropriate terminology. Ideas were culled from current examinations, and new items were constructed to cover any neglected topics. Considerable ingenuity was required to get good multiple-choice questions on some topics in order to measure understanding of function or correct procedure, rather than mere rote memory for names or phrases. Photographs and diagrams were used to add clarity and interest, and often several test items could be planned to refer to the same illustration. The tests were designed to require from one to two hours.

An experimental edition of a final achievement examination was printed and tried out on several sample classes. Answer-sheets were item-analyzed to detect errors in the scoring key, ambiguous items, and items which did not discriminate between high- and low-scoring trainees. Comments from school officers sometimes indicated other desirable revisions. The revised test was again tried out, and from the try-out scores, norms were determined and tables were constructed for converting raw scores to Navy grades. Several parallel forms of the final achievement examination were usually prepared to permit rotation of tests from one graduating class to the next. The tests, with adequate supplies of manuals, keys, and norms, were then distributed to instructor training officers for regular administration in all enlisted schools. The Test and Research Section collected test scores and final school grades for analysis of validity and for comparison of schools.

PERFORMANCE TESTS. Performance tests were intended to measure the practical skills which were not adequately measured by paper-pencil tests. To clarify the needs, a test technician worked at the training centers, familiarizing himself with the equipment, the jobs, and the local training practices, in order to decide which important jobs should be included in the testing program. In all of this he was aided by instructors in the schools.

In devising tests, three difficulties usually encountered were: (1) scarcity of equipment, (2) lack of time for testing all men, (3) objections to testing men singly rather than in teams. A major contribution of the technicians was to show instructors how to give more adequate tests in less time by using more equipment and keeping all men busy, rather than having an entire class stand by while one man was being tested. Materiel problems were met by the breakdown of equipment into subassemblies to test men on component parts of a job and by carefully scheduling the use of all equipment. To conserve time, multiple sets of equipment were used and several tests were scheduled simultaneously, so that trainees moved from written test to identification tests or performance tests, keeping busy all of the time. Time was also saved by skipping routine operations, such as the manipulation of duplicate screws. By sampling only the more significant parts of long jobs like engine assemblies, it was possible to measure the trainee's practical ability in a fraction of the time required for the entire job. The need for teams was often avoided by having a test lead up to, or start from the point where heavy parts would need to be moved.

Tests were tried out on classes and revised to shorten the time required or to improve scoring. Rating sheets which allowed considerable leeway in evaluating quality of performance were found,

in general, to be unreliable, because different instructors did not agree in grading trainee performance. For objective scoring, the proctor's check-sheets were made highly specific. This required agreement on the most acceptable procedure. From the controversies thus aroused there resulted beneficial clarification of training methods. The check-lists also made it possible to use student proctors to observe and score the test subjects, with one instructor supervising several proctors. Figure 5-xv shows a check-list for scoring the operation of a radio direction finder. Enclosed in the box near the top are the directions to the proctor concerning the condition of the equipment at the beginning of the test, the method of scoring, and (in cases where tools are involved) the list of necessary tools. Verbatim directions to the trainee are printed, followed by a list of all items to be scored, including accuracy, sequence (where it is important), tool use (in torpedoman or engineering tests), and speed (where important). The scoring is objective in that the proctor has no difficult decisions to make but merely observes what the subject does and indicates it by circling or crossing out numbers on the check-list. In some cases differential weights are assigned to items according to their importance in the training program.

When tryouts showed the tests to be satisfactory, manuals were prepared with instructions for administering and scoring the tests, and for converting the scores to Navy grades (Appendix D-1). A Bureau representative supervised the testing until the system was familiar to the instructors appointed to carry on the performance-testing program in a school.

Because they were an aid to the training program, performance tests were administered during the course as well as at its end. The frequency of such tests depended upon the nature of the training and the difficulty of setting up tests. Unit performance tests were given every four weeks in torpedoman school, every week in gunner's mate and basic engineering schools, and three times a week in landing craft school. In radio school, typing and code-reception benefited from daily testing during a large portion of the drill time.

Identification Tests. A test technician working at a training center might prepare an identification test for each week of training or to sample most of the curriculum, but in either case the attempt was to select 25 to 50 significant parts that would sample the trainees' familiarity with and understanding of a wide range of equipment. As with any multiple-choice test, the main difficulty was to get several plausible incorrect names and functions as foils for the correct names and functions. Six or eight alternatives were often listed in initial tryouts, so that the best five could be retained after item-analysis. Repeated tryouts and revisions eliminated ambiguities and

DIRECTION FINDER OPERATION TEST (Model DAE) Score

Class ..... Name ..... ☐

DIRECTIONS TO PROCTOR: Receiver turned off; VOLUME dial at "0"; BALANCER away from "0"; BAND SELECTION switch incorrect; two headphones plugged in (one for proctor); gyro repeater model within sight; calibration tables available.

Scoring: If step is done correctly, circle the number in the score column. If step is omitted or incorrect, cross out the number. Cross out time bonus scores that are not awarded.

DIRECTIONS TO TRAINEE: "DETERMINE THE TRUE BEARING OF THE STATION THAT YOU WILL FIND TRANSMITTING AT .....KC."

Time: Start..... Finish.....

Steps	*Score*
Turn power "ON" and dial light "ON"	1
Select proper tuning band	2
Before tuning	1
Switch BFO "ON"	1
Tune in station and adjust to comfortable volume	1
Tune for maximum signal strength	1
Set BALANCER at "0"	2
Unlock loop	1
Rotate loop for minimum signal	2
Adjust BALANCER for sharper null	2
Record gyro repeater reading (now, not later)	2
Record azimuth scale reading and time	2
Reduce volume very low	1
Hold BALANCER in sense position (against spring pressure)	2
Turn loop equal amounts to both sides of null point	2
Interpret signal changes correctly	2
Use Table for corrected bearing	2
Accuracy of bearing (compared to bearing obtained by instructor)	
Within 3 degrees	4
From 3 to 6 degrees off	2
Time bonus: Less than 3 minutes	6
From 3 to 4 minutes	3

Figure 5-xv. Sample Check-Sheet Used by Proctor in Administration of Direction Finder Operation Test (Model DAE)

errors. The tests were then printed in manuals with complete directions for administration (Appendix D-2).

Results of Achievement Testing

The sudden termination of the war prevented an extensive evaluation of the growing achievement testing program, but the following examples illustrate the contributions of achievement testing in (1) improving training, (2) improving service school grades, (3) improving the prediction of success in training, (4) comparing classes and instructors, and (5) measuring the achievement of schools.

IMPROVEMENT OF TRAINING. Initial tryouts of performance tests frequently revealed that men could not do the jobs they had supposedly been taught. The commonest fault of instructors was to assume that men could do a job after a lecture on theory and a demonstration. Often the equipment was not being utilized most effectively. For example, one radio school was proud of having the trainees stand watch on receivers for many hours during two weeks, in a room with five types of receivers. Yet when 25 of these men were given a standard performance test, 16 failed completely to tune in a specified station. Apparently, a large number of the trainees had been twirling dials until some station came in audibly, and instructors assumed they knew how to tune the sets. Needless to say, the instruction was soon improved. At another school the men were speedy in tuning large-ship type transmitters but had no idea how to tune the more widely used ultra-high-frequency crystal set.

At an amphibious training base, tests on fourteen "minimum essentials" revealed very poor achievement. Nearly all men failed tests on blinker receiving and compass-and-steering. On the other hand, they made very good scores on tests of markers, buoys, and flag hoist after four weeks of training. But after two additional weeks of training they made poorer scores, probably as a result of boredom with repeated lectures. The situation called for more drill on blinker and compass, and less repetition on markers, buoys, and flag hoist.

Technicians working in the schools were able to suggest ways of improving training by spending less time on lectures, breaking down equipment so that more men could use it, and giving more shop practice with supervision. Instructors, disappointed by the performance of their trainees, thought of many ways to improve training. The net result was an improvement of performance test scores in later classes. The example shown in Figure 6-xv illustrates the situation in a torpedoman school. Since different parallel tests were tried out in successive classes, several weeks elapsed between the first and

second administration of a particular test. The men did not know what jobs would be tested, so the improvement in performance indicated better training, or more practice by the trainees.

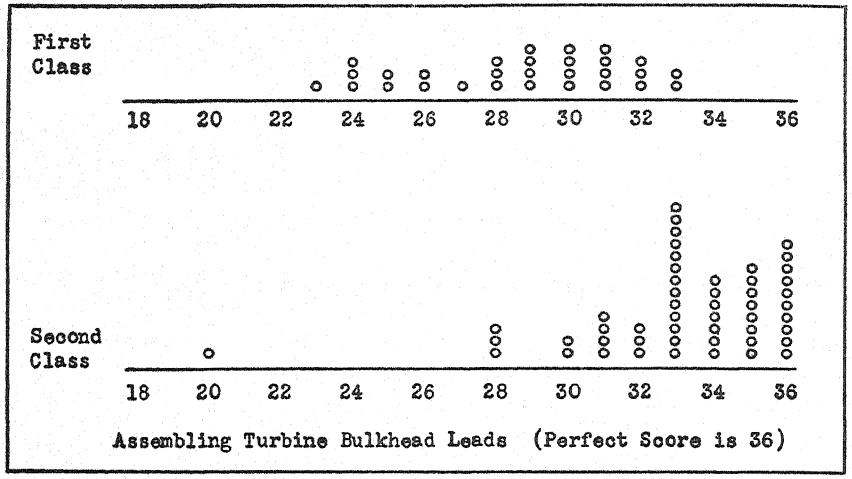


Figure 6-xv. Improvement of scores in Torpedoman Performance Test (assembling turbine bulkhead leads) in successive classes at a Torpedoman School.

Likewise, identification test scores improved in successive classes, as illustrated by Figure 7-xv for the Small Arms Test at a gunner's mate school. Similar examples for each school suggest that achievement testing aids training by revealing deficiencies of instruction

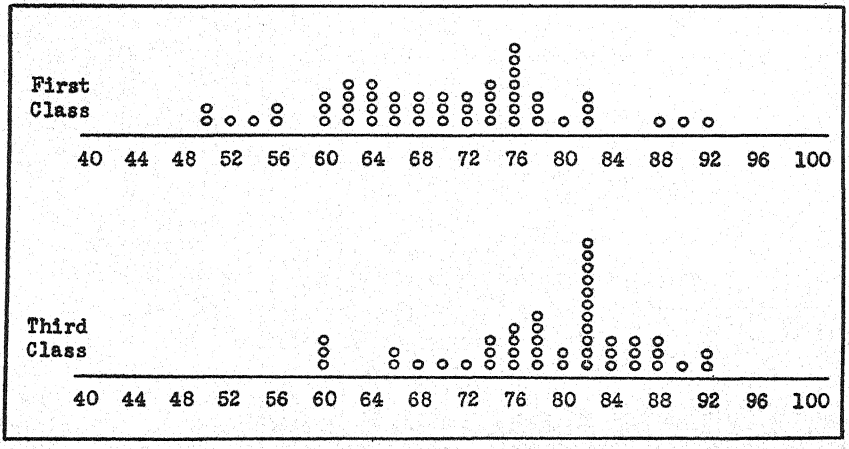


Figure 7-xv. Improvement of scores in Small Arms Identification Test in successive classes at a Gunner's Mate School.

and by encouraging trainees to practice the jobs on which they may be tested.

IMPROVEMENT OF SCHOOL GRADING SYSTEMS. It was frequently observed that practical ability, an important objective of enlisted training, was not adequately reflected in final school grades. The schools' usual written tests neglected this aspect of training, and shop ratings of practical ability were often too uniform to make an effective contribution to final grades. One service of standardized achievement examinations and performance tests was to give more appropriate weighting to all aspects of the curriculum. Consider,

TABLE 1-XV. The relative contribution of each part-grade to the total variance of the composite final grade for graduates of two classes in a Basic Engineering School

	Statistics of Part-Grades						Relative Contribution[1] of Each Part to Total Variance	
	Class I N = 350			Class II N = 340				
Part-Grades	M	σ	r_{jT}	M	σ	r_{jT}	Class I	Class II
Mathematics	83.6	7.7	.86	83.4	8.5	.88	.61	.62
Mechanical Drawing	89.1	4.1	.74	87.7	4.1	.72	.28	.25
Shop	84.0	2.5	.48	83.5	2.6	.60	.11	.13

The formula[1] used was $K_j = \frac{W_j \sigma_j r_{jT}}{\sum W_j \sigma_j r_{jT}}$

Where K_j = relative contribution of part j to the total variance.
W_j = formal weight of part j (in this case $W_j = 1.00$).
σ_j = standard deviation of scores on part j.
r_{jT} = correlation of part j with composite total T.
X_j = score on part j (j being successively the three part grades).
$T = W_j X_j$.

[1] This is a modification of a formula developed by Richardson. See Richardson, M. W., "Combination of Measures," in Horst, Paul, *The Prediction of Personal Adjustment*, Supplementary Study D, Bulletin 48, Social Science Research Council, New York, 1941, p. 383.

for example, a basic engineering school where four-sevenths of the time was spent in shop work and three-sevenths of the time in classroom work, including mathematics, mechanical drawing, and shop theory. A survey of grades of 690 men in two graduating classes (Table 1-xv) showed that the relatively unimportant mathematics tests contributed much more to final grades than did the important shop work. The last two columns of Table 1-xv indicate that shop grades contribute to final grades less than one-fifth as much as do mathematics grades, although they should contribute four times as much in terms of time spent. This very small contribution of shop grades might be expected in view of their small standard deviation,

only 2.5 as compared to a standard deviation of about 8.0 for mathematics grades. To correct this situation it was necessary to weight mathematics less in the final average, and to increase the contribution of shop grades.

A large part of the shop-work grade of the first four weeks came from ratings of the hand-tool work of the trainees in making "samplers" to specifications. When 30 such samplers were graded by four different instructors, by the usual method with combination squares, there was great disagreement among the grades assigned to the same samplers. Their ratings intercorrelated from —.11 to .55. A set of taper gages and caliper gages was devised, with scales for five points of deviation on either side of specifications (Figure 4-xv). When these gages were used in scoring samplers, the ratings of two different instructors correlated .93 on the second week sampler, and .96 on the third week sampler. This increase in objectivity of ratings was accompanied by an increased spread of grades, so that different levels of performance were more effectively discriminated.

For the last four weeks of basic engineering school, identification tests and performance tests were devised to measure trainees' familiarity with naval machinery and their practical ability in assembling and operating it. Not only did this aid instruction but it also provided more dependable shop grades with enough dispersion to contribute appropriately to final school grades.

Less striking changes in grading were made in most schools where standard tests were introduced, but the changes usually involved more stress on practical ability. Although it had been assumed that written tests sufficed to indicate what a man had learned in a service school, the evidence showed that performance tests and improved shop grades were not closely correlated with written test grades. During test tryouts in a gunner's mate school, performance test scores correlated from .14 to .35 with written tests, and only slightly higher with final grades, which were based largely on written tests. In a torpedoman school, where shop grading was quite good, test tryouts showed that, on the average, three sample performance tests correlated .63 with final grades but only .38 with the multiple-choice final examination. Obviously the contribution of shop work and performance tests differed from that of the written tests; and with more adequate measures, practical ability could be given even more weight in final grades.

Improvement of Prediction of Training Success. The selection of prospective school trainees has been quite successful, using service school grades to evaluate the predictions made from the Basic Test Battery. But it appears that prediction can be even more successful if the school grades adequately reflect the special abilities desired

for various Navy ratings. One contribution of standard achievement examinations and tests is to make school grades more indicative of these special abilities. Consider, for example, the previously mentioned basic engineering school, for which two studies of Basic Test Battery scores in relation to grades were made about a year apart. Figure 8-xv shows the correlations, corrected for restriction in range of scores, between final grades and scores on the six tests of the Basic Test Battery. In the first study, the Arithmetical Reasoning Test was the best predictor of success in basic engineering school, and the Mechanical Knowledge Test (Mechanical Score) the poorest. But a year later, after achievement tests and shop ratings had put greater emphasis on practical ability, the Mechanical Knowledge Test became the best predictor of basic engineering success, and the

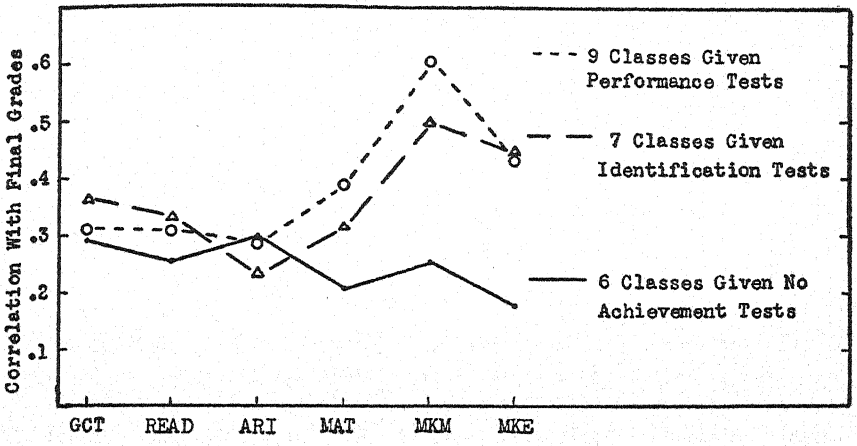


Figure 8-xv. Prediction of success in Basic Engineering School by use of the Basic Test Battery (before and after introduction of Achievement Testing Program).

verbal and numerical aptitudes became less important, as seems appropriate.

An equally striking change in torpedoman school is illustrated in Figure 9-xv. The solid line, representing six classes that took no standardized achievement examinations, shows that none of the basic battery tests had been very useful in predicting success. The dashed line represents seven contemporary classes in the same school on which identification tests were tried out. Although test results counted little toward final grades, the indirect influence of the tests was to put more emphasis on practical ability, so that the students' final grades more nearly conformed to their mechanical aptitudes. The dotted line in Figure 9-xv represents nine classes on which sample performance tests were tried out. These scores were included

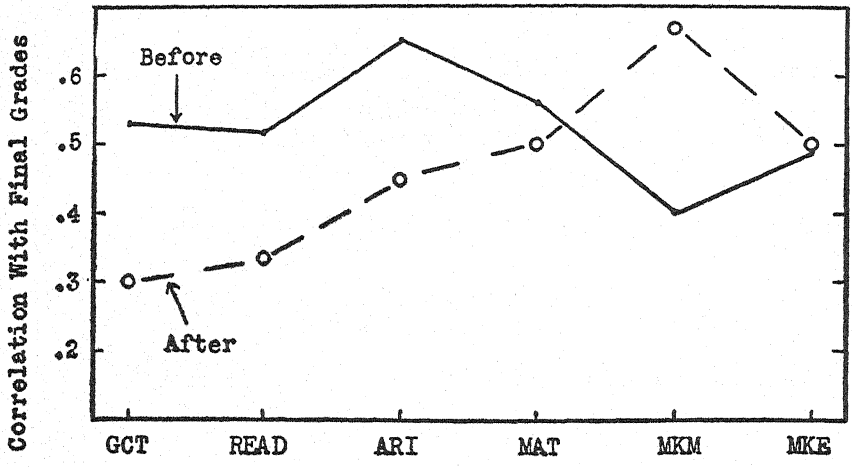


Figure 9-xv. Prediction of success in Torpedoman School by use of the Basic Test Battery (before and after introduction of Achievement Testing Program).

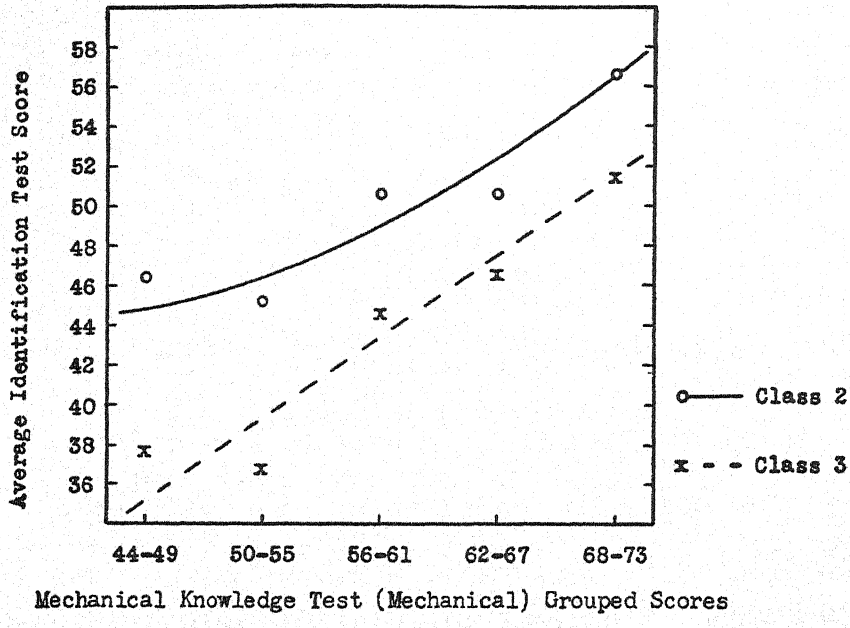


Figure 10-xv. Comparison of performance on Torpedoman Identification Test of two classes. Distributions of grouped scores on the Mechanical Knowledge Test (Mechanical Scores) are given to permit comparison of classes.

in the final average, so that in this case the final grades correlated even more highly with the Mechanical Knowledge Test (Mechanical Score). These results show that wider use of practical achievement tests would make it easier to predict success in torpedoman school from the Mechanical Score on the Mechanical Knowledge Test.

COMPARISON OF CLASSES OR INSTRUCTORS. Comparison of classes is nearly impossible when instructors make up their own examinations. By providing a dependable basis for comparison, standardized achievement examinations are an aid in assigning grades that are comparable from one class to another. The tests can also be useful in spotting weak instruction. As an example, in a torpedoman school, Class 3 made much lower scores on an identification test than did Class 2 which had taken the same test one week earlier. The school officer suspected poor instruction, but the instructors insisted

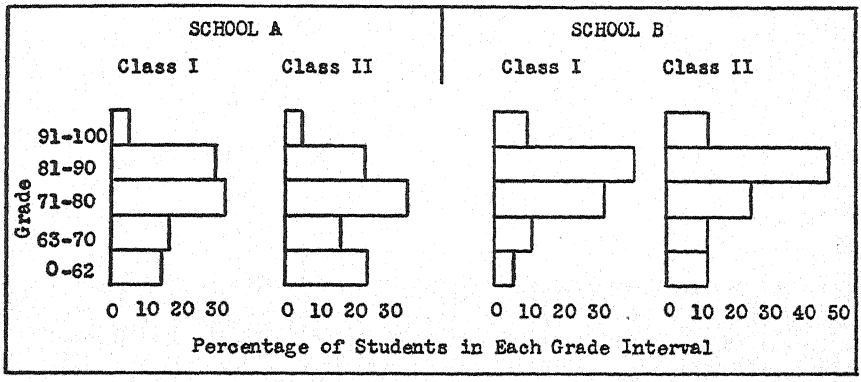


Figure 11-xv. Comparison of classes in two Torpedoman Schools as to performance on Torpedoman Final Achievement Examination (N = approximately 240 in each class).

that Class 3 was just stupid. As the Mechanical Knowledge Test (Mechanical Score) was known to be the best predictor of identification test scores, the plot of Figure 10-xv was made, to clarify the issue.

The plot is simplified to show merely the *average* Identification Test score for men in each score interval of the Mechanical Knowledge Test (Mechanical Score). The range of aptitude scores is the same in both classes, yet the regression line of best fit shows that the men of a given level of aptitude on the Mechanical Knowledge Test (Mechanical Score) make higher scores on the achievement test if they are in Class 2. Evidently there was poorer training in Class 3. The curved plot for Class 2 shows how, by better instruction, the poorest students can be brought up to a comparatively good level of achievement.

When the same standardized tests are given in different schools of the same type, it becomes possible to compare classes from one school to another. Figure 11-xv illustrates such a comparison in terms of a printed final achievement examination. Grades were reported to the Test and Research Section in five groups, with grades from 91 to 100 in the highest group and unsatisfactory grades (62 or

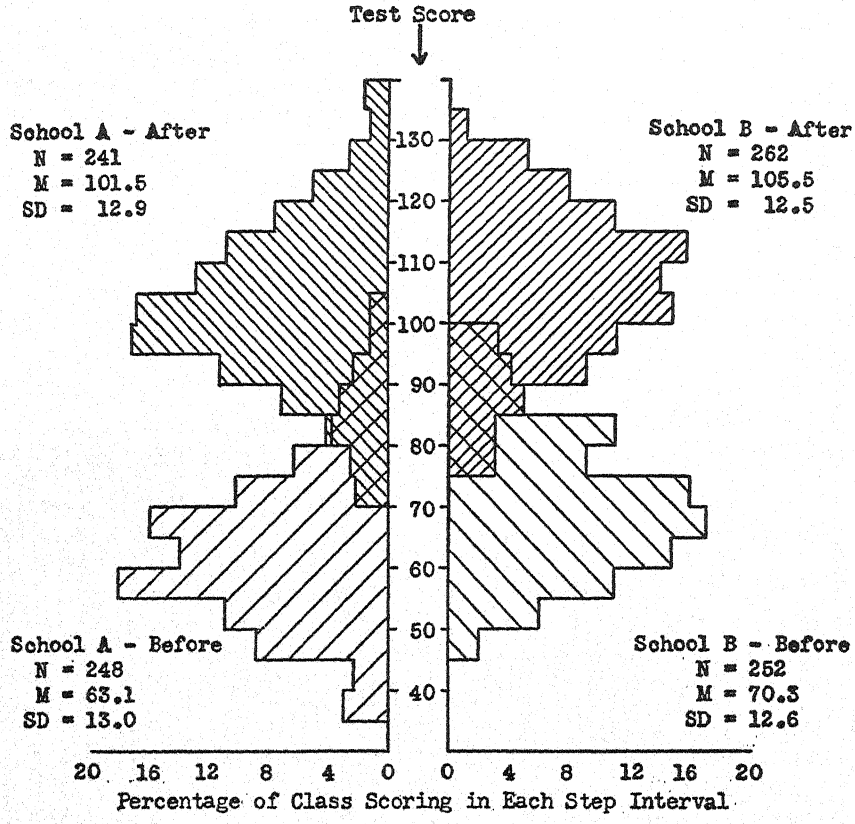


Figure 12-xv. Comparison of performance of classes in two Basic Engineering Schools before and after training, measured by a standardized Final Achievement Examination.

below) in the lowest. The bar graphs show what percentage of trainees in each class received grades in each interval. For example, in Class 1 of School A, 5 per cent of the men had achievement test grades from 91 to 100, 30 per cent had grades from 81 to 90, etc. Comparison of the bar graphs shows no great difference between successive classes within a school, but the classes in School B are consistently superior in achievement to the classes in School A. In order

to account for such differences, or to remedy them, specific weaknesses of training must be analyzed.

MEASURING THE ACHIEVEMENT OF SCHOOLS. Achievement tests can be used to show how much a class has gained from its training, to answer those who consider school training to be useless. To measure the gain, classes in two basic engineering schools were tested before and after their training course, with parallel forms of the printed final achievement examination. Form II was given to the beginning classes, and Form I was given after the course was completed. The two forms were constructed with very similar items, and had been shown to give practically identical scores. In Figure 12-xv the results of the study are shown, with School A on the left and School B on the right. The crosshatched areas indicate the overlap of scores between men who scored highest on the first examination and the men who scored lowest on the later examination. It is clear that there is a great increase of knowledge after instruction at either school. It is also apparent that the trainees at School B scored slightly higher than those at School A, both before and after training. Some observers had considered the training at School B to be much superior, but from this test there is no evidence that trainees *gained* more at School B than at School A.

If an achievement test is subdivided by topics, a more adequate analysis can be made of strong and weak aspects of training in various schools. This is illustrated in Table 2-xv by the final achievement examination for recruit training. The 300 items of the test were divided into nine topical sections, listed in the first column of the table. For each part-score the means and standard deviations are reported for four different training centers. If there were any outstanding deficiencies at one training center, they should become apparent from this detailed comparison of test scores. But the only evident trend in Table 2-xv is that trainees from Center B show superior achievement in total scores and in all sections except Gas Warfare Defense. If it were known that recruits of similar ability were sent to all centers, one could conclude that the training at Center B was definitely superior. By comparing scores of later classes with those reported in Table 2-xv, school officers could observe whether or not there was improvement of training on specific topics.

IMPROVEMENT OF MOTIVATION. Although there are no statistical data on morale, it was consistently observed that the development and use of practical shop tests stimulated the interest of instructors and students. Most instructors, having felt the need for better evaluation of practical ability, were pleased to find that performance tests were actually feasible. And when students performed poorly, the instructors were quick to try to improve training. The students were

TABLE 2-XV. Achievement of recruits at four Naval Training Centers as indicated by means and standard deviations on the Recruit Training Final Achievement Examination, Form 1

Recruit Training Final Achievement Examination Section	Number of Items	Center A N = 589		Center B N = 858		Center C N = 644		Center D N = 617	
		M	σ	M	σ	M	σ	M	σ
Naval Orientation	40	24.22	4.90	27.93	4.63	26.30	4.65	24.44	4.92
First Aid and Personal Hygiene	25	17.13	4.82	20.78	3.76	18.87	4.63	17.20	5.78
Seamanship	80	38.26	9.32	50.10	9.80	40.54	10.03	42.38	10.92
Recognition	30	18.54	5.11	23.10	4.57	19.78	5.23	20.00	5.40
Lookout	25	17.10	3.89	20.83	2.81	17.84	3.93	17.60	4.06
Telephone Talker	25	14.14	4.21	17.67	3.54	16.30	4.09	16.62	4.28
Ordnance and Gunnery	25	13.04	2.94	14.51	2.70	13.75	2.94	13.56	3.22
Gas Warfare Defense	25	15.61	3.83	15.09	2.94	14.14	3.41	13.50	3.73
Fire Fighting	25	11.35	3.78	16.55	4.06	14.32	4.24	13.70	5.21
Total Test	300	169.56	32.93	206.91	29.81	182.08	32.58	178.99	37.28

also impressed by the importance of performance tests. Even the least motivated of them took pride in making good performance records, although they might not take the written tests seriously. Identification tests stimulated interest in learning about all kinds of equipment. After a test the students would discuss parts among themselves or with instructors, and in at least one school the comparison of parts from different torpedoes became a favorite pastime. It seems likely that improvement of test scores resulted from increased motivation of students as well as from better instruction.

Summary

The experience gained during the war in measuring achievement in enlisted training programs emphasizes the need to make test items which measure functional knowledge. Probably naval instructors are more insistent than many in civilian education on the importance of practicality in a given situation. Hence, in order to develop tests which instructors will consider adequate and which will gain the respect of instructors and students alike, it is essential that the items reflect the functional aspects of the training program. Such test items should deal with actual procedures of sending or interpreting signals, overhauling or adjusting equipment, analyzing casualties, or operating and securing equipment. Whenever possible, the tests should require the handling of real or model equipment, rather than talking about it. When this is not feasible, extensive use should be made of diagrams and pictures of equipment.

The program also gave convincing evidence that an important outcome of the use of standard achievement tests may be to stimulate better training in enlisted schools. By testing jobs that were fairly similar to the jobs the trainee might encounter in later billets, it was often demonstrated that training was not fulfilling certain of its purposes. Instructors did not mean to be impractical, but they simply did not realize their impracticality until objective tests gave the evidence. After standardized tests were introduced, there was much more uniformity of training among parallel schools. Ambiguities and differences of opinion showed up clearly when one tried to construct objective tests. And in the process of developing satisfactory tests, shop procedures were often improved and procedures of operation and overhaul were clarified, to the enlightenment of both instructors and students.

Each measure of achievement serves a somewhat different function. Product-rating gages help to make shop grades more dependable. Identification tests motivate trainees to understand or recognize a wide variety of equipment. When broad coverage of theory

and procedure is desired, the multiple-choice paper-pencil test is most efficient. Performance tests help to make sure that men can *do* jobs. Such tests stimulate a great deal of practical training even if only a few sample tests are given to each man.

In general, the achievement testing program applied the best already-known methods to the rather new and unique situation in each enlisted school. Although much ingenuity was exercised in developing tests of military operations, it would be hard to demonstrate the invention of any techniques that had not been tried previously in industry or educational institutions. The program did provide convincing evidence that objective methods of evaluation can be applied to practically any job, however complex or mechanical it may be.

CHAPTER XVI

ACHIEVEMENT EXAMINATIONS FOR OFFICER SCHOOLS

THE problems of examination in the Navy's schools for the training of officers were essentially the same as those in the enlisted training schools with the additional complications implied by (1) the wider range of subject matter to be covered and (2) the greater complexity of the concepts, principles, and procedures to be measured. The need for assistance to the schools in the development of objective and valid measures of achievement was early recognized by the Test and Research Section, but personnel were not available to accomplish the task. To an even greater extent than was true in the enlisted schools, the officer schools were left on their own in the development of curricula and teaching materials. As schools of the same general type were established in new locations, they tended to take over the practices of those which had been established earlier, adapting their programs to the facilities and equipment available in the new location and modifying the content and organization of the instructional program in the light of the experiences and preferences of the administrative and instructional staffs. The necessity of shifting a small percentage of the trainees from one school to another forcefully directed the attention of the school authorities to these differences, and this in turn led to a healthy interchange of opinion as to the desirability of continuing or instituting certain units of instruction and as to the relative value of different sequences and modes of presentation. The outcome of such interchange was a demand, originating in the schools, that standardized curricula be provided and that some means other than the naked opinion of the training staffs be provided to estimate the relative effectiveness of the various programs.

The staff of the typical wartime officer training school was made up of officers drawn from three sources: (1) retired officers and regular line officers placed in a limited duty status, (2) reserve officers who were procured for teaching duty from positions in civilian schools and colleges, and (3) recent graduates of the training schools selected from the highest ten per cent of the graduating groups and held over to guide and instruct the new incoming groups of trainees. The first and third of these groups were ordinarily unfamiliar with modern training methods, the second and third tended to be unfamiliar with the Navy, and even the instructors drawn from civilian schools found some difficulty in adapting themselves to the accelerated pace and the emphasis on practice and procedure rather than

theory and principle, which were characteristic of the training programs. But all recognized the critical importance of their assigned mission and undertook to do the job to the best of their collective abilities within the limitations of time, facilities, and equipment available.

In this climate of recognized need, efforts of the Bureau of Naval Personnel's Training Activity to develop uniform curricula, to provide suitable uniform equipment and training aids for implementation of the curricula, and to introduce objective and reliable measures of achievement were assured the ready cooperation of the schools.

Procedures Used in the Development of Examinations

While the procedures used in the development of examinations for officer training programs were necessarily varied in order to reflect the conditions, curricular content, and the number of different types of schools, there was a uniform general pattern of test development. The first step in this procedure was to attempt a definition of the objectives which the school's program was designed to attain in terms of the knowledges and skills which were to be expected of its graduates. This necessitated a review of the curriculum for the school, supplemented by examination of the specific courses of study and the schedules of class instruction and drill in effect at the various installations. This study of the objectives and actual curriculum culminated in the preparation of an examination outline which was submitted to the school or schools for criticisms and suggestions for revision.

Following the development of an examination outline, or sometimes concurrent with it, the work of assembling and preparing examination materials was undertaken. Usually the schools for which an examination was being prepared were asked to submit copies of the tests which they had already used with current and previous classes, sometimes with samples of the answer sheets or examination papers for statistical analysis. In addition to review of the testing materials submitted by the schools themselves, the officers working on an examination studied the texts, reference materials, and recommended films and prepared examination items based upon these materials. From this stock of examination materials, tentative forms of the proposed examination were compiled, allocating the items to the headings of the examination outline in proportion to the relative weights assigned to these headings. The weights assigned were based on the amount of time devoted to the treatment of each topic in the course of study and schedules, tem-

pered by the judgment of the instructors as to the relative importance of the unit or topic.

The tentative forms of the examination were submitted for criticism to various experts either in the schools, or in more advanced specialized schools, or in the office or bureau of the Navy Department having cognizance over the content of the material. These fundamental questions affecting the validity of the tests were asked: (1) Is the information given in the item setting, and the keyed response, correct? (2) Are the decoy answers such that a person without the correct knowledge would be equally likely to choose one of them? (3) Does the item as a whole require the application of knowledge or judgment that the graduate of this school should be expected to have? (4) Has the information or concept necessary to give the correct answer to this item actually been taught in the school? (5) Does the whole examination provide a reasonably adequate representative sample of the desired outcomes of the course?

In addition to the review of the examination materials to eliminate error and assure validity as to content, each item was scrutinized to discover extrinsic clues in the structure of the items, e.g. having the correct response indicated by its greater length or specificity, or by the absence of grammatical or logical sequence in the distractors. Items were also checked to avoid ambiguity and to assure the use of standard or conventional terminology.

The next step was the preparation of experimental forms of the examinations. These were either in mimeographed or photo-offset reproduction, depending upon the size of the group to be used in the experimental administration and upon the nature of the test content, especially taking into account the number and character of illustrations used and the need for precise representation of figures and diagrams. An officer of the Test and Research Section was usually assigned to supervise the administration of the experimental forms in order to obtain first-hand data on such matters as (1) the amount of time that should be allowed for the examination period, (2) questions asked by the student-officers which should be obviated by revision of the directions, (3) specific items which the student-officers considered ambiguous, (4) comments of the student-officers as to the adequacy and fairness of the test, and (5) further criticism and correction of the test content by the instructors who served as proctors. In some instances junior instructors who had not previously seen the tentative forms took the examination along with the students in order to supply a further check on correctness of terminology and content and absence of ambiguity.

The results from the experimental administration of a test were subjected to statistical analysis to determine the reliability of the

test, and of the parts if separate part scores were indicated, to obtain the difficulty index of each item and to determine the discriminative capacity of each item. Frequency distributions of the scores were made and the means and standard deviations computed. Where two or more forms of the examination were being prepared simultaneously, these measures were used in determining the equivalence of the forms. The indices of difficulty and discriminative capacity were used in selecting items for inclusion in the final forms and as a means of spotting items which, despite all the precautions taken, proved to contain extrinsic clues or to be incorrect or ambiguous.

After final forms of an examination were prepared, a further experimental administration was conducted in order to establish tentative norms. Generally the officer training schools adhered to the Navy's "4.0" grading system in which a grade of 2.5 is the lowest passing grade, 3.0 is average, and 4.0 the highest degree of attainment. Procedures for translating the raw scores on the examinations into this type of grade varied. In some cases an arbitrary decision was made as to what raw score on the examination should be equated to 2.5 representing the lowest passing mark. Similarly a raw score somewhere between the highest score attained and the highest possible score on the test was equated to 4.0. Intermediate scores were then translated to grades on a linear scale and grades lower than passing were computed by extrapolation. Another procedure equated the mean of the raw scores to 3.0 and fixed the raw score to be equated to 2.5 in terms of a multiple of the standard deviation of the distribution. For example, on a test having a mean of 126.3 and a standard deviation of 12.21, the conversion scale was determined by equating 126 to 3.0 and finding the raw score equivalent for 2.5 by subtracting 1.75 times 12.21 or 22.37 from the mean score. This yielded 103.93 or 104 as the score to be equated to 2.5.

Types of Examinations Developed

The examinations prepared by the Test and Research Section in cooperation with the officer training schools were "paper and pencil" tests. Development of performance tests for officer schools was contemplated but had not been undertaken when Japan capitulated.

While the examinations uniformly consisted of items set up in multiple choice form, the tendency to test only on fragmentary bits of information was avoided by introducing problem situations which were used as the basis for a series of items covering several critical steps or phases of the same process. This may perhaps be best shown by illustrations drawn from materials similar to those included in the examinations.

The following example is taken from material for the Damage Control section of the examination for the reserve midshipmen's schools.

The *U.S.S. Flushing* has the following characteristics: length 560′; beam 55′; draft 16′5″; displacement 8,000 tons; waterline area 22,400 square feet; moment to alter trim 1″ = 360 foot tons; moment to heel 1° = 280 ton feet; GM = 3.75′; BM = 12.60′; KG = 15.80′. She is in condition II when a torpedo hit is suffered amidships. Eighty tons of flooding water are taken aboard in a waterline compartment on the second deck. The compartment, whose inboard bulkhead is 7½′ from the C.L. of the ship, is 25′ long, 20′ wide, and 7′ high.

The following series of problems, items 1 through 6, is based on the above information:

1. The transverse free surface effect causes a reduction in GM of
   1. 0.06′
   2. 0.10′
   3. 0.55′
   4. 0.71′
   5. 1.80′
2. The free communication effect of the flooding water reduces GM by about
   1. 0.10′
   2. 0.16′
   3. 0.24′
   4. 0.33′
   5. 0.55′
3. The tons-per-inch immersion of the *Flushing* in sea water is
   1. 19.1
   2. 22.2
   3. 28.6
   4. 35.8
   5. 53.3
4. After the *Flushing* takes on 105 tons of fuel oil from an oiler at sea, her mean draft will be increased by about
   1. 1″
   2. 2″
   3. 3″
   4. 4″
   5. 5″
5. The *Flushing*, after repair of all previous damage, takes a torpedo hit aft which puts 80′ of the waterline in free communication with the sea. The approximate GM for the ship in the damaged condition is
   1. 1.80′
   2. 1.95′
   3. 3.21′
   4. 3.43′
   5. 5.55′
6. The *Flushing* is 6″ down by the head. There are 60 tons of fuel oil in a forward tank, which may be pumped into one of four tanks, A, B, C, or D, at points 30′, 50′, 90′, and 180′ farther aft, respectively. To eliminate the 6″ trim by the bow and cause her to ride with a *slight* drag, the oil must be pumped into tank
   1. A.
   2. B.
   3. C.
   4. D.
   5. None of these tanks.

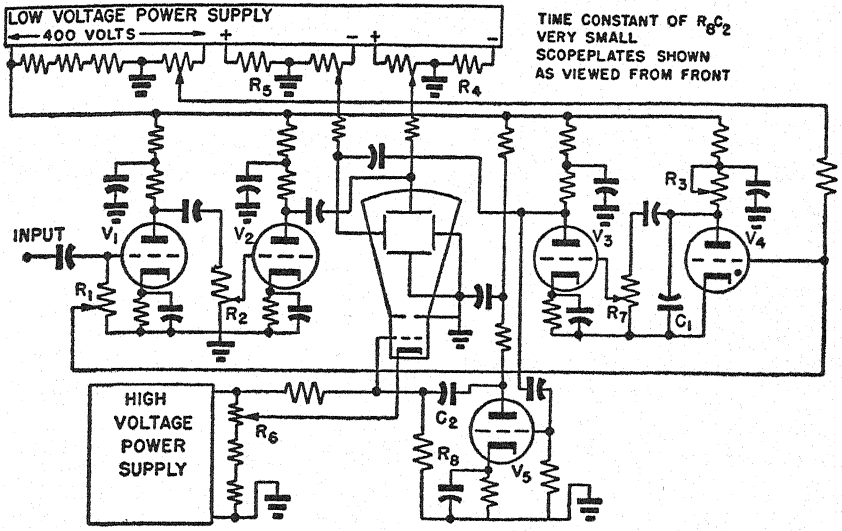
Another example of a series of test items requiring interpretation of technical data in a situation presented by diagram is drawn from the pre-radar field (Figure 1-xvi).

A device which might be considered somewhat as an innovation in objective testing was used in the navigation section of the examination for the reserve midshipmen's schools. In this examination, the students were provided with a leaflet of excerpts from various publications such as the Nautical Almanac, Tide Tables, and Current Tables, and were instructed to bring with them certain basic reference materials. They were also provided with a chart on which to work out a series of problems simulating some of the elements of a navigator's day's work at sea. The first section of this test dealt with simple informational material. The second section required the use of the references to solve simple problems in nautical astronomy. The third section required the construction of a continuous navigational plot from the data given. The latter two sections were weighted in scoring to compensate for the additional time required in locating data in the references, computation, and plotting. The ninety minutes allowed for 33 items comprising these sections was found to be none too much for the average midshipman.

Use of the Examinations

The examinations were designed to serve as supplementary measures of individual student achievement in the officer training schools. Those prepared for the pre-radar schools and for the tactical radar school served that purpose. The examinations prepared for the reserve midshipmen's schools had only reached the experimental administration stage when the schools were closed. But in all cases, reports of the examination results were made to the schools concerned for the information and use of the instructional staffs. To some extent, the very fact that a standardized examination was to be given caused instructors to adhere more closely to the curricula prescribed for their respective schools and courses and had the further effect of motivating the students to study for retention and application to situations different from those which had been specifically covered in their instruction. In the case of the pre-radar and midshipmen's schools, where there were several schools of the same type using the examinations, it was assumed that there would be comparisons made from school to school and some degree of rivalry was apparent even while the examinations were in the experimental stage.

In addition to the usefulness of the examinations as motivating devices in the training program, they served three important func-

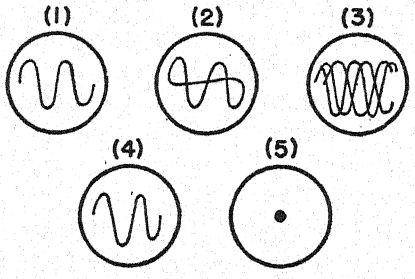


Items 1-4 are based upon the above diagram. After examining the diagram, proceed to the items below.

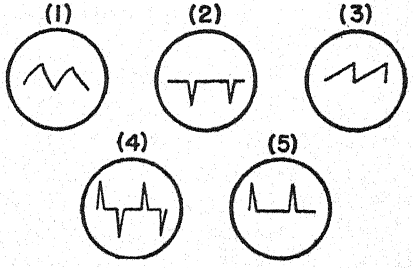
1. In the above diagram, the frequency of the sweep can be changed by varying:

   1. R_1
   2. R_2
   3. R_3
   4. R_4
   5. R_5

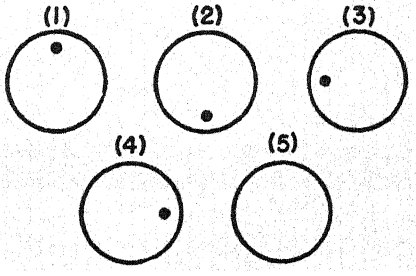
2. In the above diagram, if +400 V. decreases to +150 V. and C_1 is replaced by a smaller capacitor, a sweep frequency of 800 c.p.s. is then possible. When the sweep is again synchronized so that 2 cycles are shown on the screen, the pattern will appear as:



3. In the above diagram, the oscilloscope is operating normally. If the Y input to a Dumont 208 oscilloscope with a properly synchronized sweep is placed between the grid and cathode of the cathode-ray tube, the pattern on the screen of the 208 tube will be:



4. In the above diagram, assume that the input signal is removed and the sweep adjusted until the screen shows centered spot. If R_5 is moved to the right, the screen will appear as:



tions in developing greater standardization of instruction and grading. The data presented below will illustrate the contributions made by certain of the examinations in bringing about (1) standardization of instruction in cases where the schools were established to serve the same training purposes, (2) a means of comparing the performance of trainees in different schools or in different courses having the same general objectives, and (3) an improvement in grading and marking.

STANDARDIZATION OF INSTRUCTION. Training designed to prepare men for the Naval Training School (Radar) was established at three pre-radar schools. The length of the pre-radar course was approximately four months. It was found that considerable lack of uniformity existed in the preparatory training of men entering the radar school. Those selected for training at Pre-Radar Schools A and C were generally men with an engineering or electrical background

TABLE 1-XVI. Means and standard deviations for Pre-Radar Final Achievement Examination—first administration of experimental forms.

Form Number	Number of Items	Reliability Coefficients (K-R)[1]	Pre-Radar School A			Pre-Radar School B			Pre-Radar School C		
			N	M	σ	N	M	σ	N	M	σ
1-x	70	.68	60	46.6	6.0	56	47.1	6.7	28	42.8	7.8
2-x	110	.85	59	80.4	10.8	71	76.9	11.9	43	71.6	9.9
3-x	105	.82	43	73.5	10.1	64	68.9	10.5	44	65.4	11.2
4-x	110	.87	38	73.3	12.7	55	69.4	14.7	36	74.6	10.9

[1] Estimated by Kuder-Richardson Formula No. 21.

while those assigned to School B had simply a scientific background. School A's staff of instructors were all naval officers especially selected for the job; those at Schools B and C were, for the most part, civilians. Most important, however, at the outset, was the absence of a standardized curriculum and standard measures of achievement for the pre-radar schools. A standard curriculum was put in operation in August 1944 and the development of standardized achievement examinations was initiated in October 1944.

The data in Table 1-XVI show a marked trend toward equalization of mean scores for the classes graduating from the three pre-radar schools. These data are drawn from results of the first administrations of the experimental forms of the Pre-Radar Final Achievement Examinations. In reading this table, comparisons should not be made between means on different forms of the test, since the number of items per form and the distribution of items between the two parts of each form were not the same.

COMPARING THE PERFORMANCE OF TRAINEES IN DIFFERENT SCHOOLS OR COURSES. The use of a final achievement examination in comparing the performance of individuals undergoing different courses of instruction is illustrated by the data in Table 2-XVI. These data are drawn from the results of administration of experimental Forms 1 and 2 of the CIC Final Achievement Examination to three pairs of groups at the Tactical Radar School. In each pair, the group marked with an L (5L, 6L, 7L) was made up of student officers who were assigned to the school directly from midshipmen's or indoctrination schools. These groups received eight weeks of instruction. The groups marked with an S (6S, 7S, 8S) were sea-experienced officers who took a four-week course of instruction. As shown by the data in the table, the student officers taking the longer course performed better, in terms of scores on the objective examination, than did the sea-experienced officers who took the shorter course. While these data do not provide an adequate basis to appraise the relative merits of the two courses, it can be seen that the longer course of instruction resulted in a higher level of achievement than could be reached in a short course of four weeks, despite the fact that the latter groups had the initial advantage of sea experience.

A further example of the use of the examinations to compare the performance of students in different schools is presented in Table 3-XVI. The data in this table are taken from the results of the experimental administration of the Reserve Midshipmen's Schools' (Deck) Standardized Examination in two of the schools. In one of these schools there were separate departments of Seamanship and of Engineering and Damage Control. In the other, instruction in Engineering and Damage Control was carried on by the Seamanship instructors during the last few weeks of the course.

From these data it will readily be seen that, on the Engineering and Damage Control section, the performance of students in School B was markedly inferior to that of the group in School A. School A's median score was higher than the scores obtained by 94 per cent of the students in School B, and fewer than 10 per cent of the School A group fell below the School B median. The difference between the two means is significant above the one per cent level. While the difference between the groups on the Seamanship section was not as great, it was more striking in view of the fact that it had not been anticipated. School A's median score on this test was higher than School B's 70th percentile, while only 22 per cent of the group in School A fell below School B's median. Here too, the difference between the two means is significant above the one per cent level. Thus it appears that the combination of Engineering and Damage Control with Seamanship as handled by School B not only resulted

TABLE 2-XVI. Performance on the CIC Final Achievement Examination of student officers receiving eight weeks of instruction and sea-experienced officers receiving four weeks of instruction at Tactical Radar School

Score Interval	Distribution of Scores on CIC Final Achievement Examination											
	Form X-1								Form X-2			
	Class 5L		Class 6S Sea Experienced		Class 6L		Class 7S Sea Experienced		Class 7L		Class 8S Sea Experienced	
	Number of Cases	Cumulative Percentage	Number of Cases	Cumulative Percentage	Number of Cases	Cumulative Percentage	Number of Cases	Cumulative Percentage	Number of Cases	Cumulative Percentage	Number of Cases	Cumulative Percentage
250-4					1	100						
245-9												
240-4												
235-9	4	100			1	99						
230-4	3	97			1	98						
225-9	4	95			9	97	1	100	1	100		
220-4	6	92			13	90						
215-9	8	87			15	80	1	96	1	99		
210-4	11	82	1	100	17	68	1	93	6	98		
205-9	11	74	1	97	12	54	2	89	4	94		
200-4	10	66	3	93	7	44	2	81	10	91		
195-9	15	57	2	83	14	39	3	74	10	83	1	100
190-4	13	46	1	76	11	27	1	63	11	75		
185-9	13	37	3	72	6	19	2	59	14	66	2	96
180-4	8	27	5	62	6	14	1	52	19	55	3	89

175-9	11	21	3	45	1	9	4	48	10	41	4	78
170-4	5	13	4	34	4	8			13	33	1	63
165-9	7	9	3	21	1	5	2	33	3	23	5	59
160-4	1	4	1	10			2	26	3	20		
155-9	1	3			2	4	2	19	9	18	1	41
150-4	1	2			3	2			4	11	2	37
145-9			2	7			1	11	3	8	2	30
140-4	2	1					1	7	3	5	2	22
135-9							1	4	1	3	1	15
130-4									2	2		
125-9									1	1		
120-4											2	11
115-9												
110-4												
105-9												
100-4											1	4
Q_3	211.1		193.7		217.3		200.6		194.9		179.9	
Md	196.7		181.5		207.9		182.5		183.2		167.5	
Q_1	183.4		171.6		193.6		164.4		171.2		146.8	
Mean	196.40		181.31		204.0		182.20		180.71		161.25	
σ	19.80		15.90		18.75		23.20		19.75		22.50	

in poorer achievement in the field of Engineering and Damage Control but also detracted from student achievement as measured by the examination at the end of the course in Seamanship. This

TABLE 3-XVI. Performance of students in two Reserve Midshipmen's Schools on Parts 1 and 2 of the Reserve Midshipmen's Schools' (Deck) Standardized Examination

Part 1—Seamanship					Part 2—Engineering and Damage Control				
	School A		School B			School A		School B	
Score Interval	Number of Cases	Cumulative Percentage	Number of Cases	Cumulative Percentage	Score Interval	Number of Cases	Cumulative Percentage	Number of Cases	Cumulative Percentage
160-63			1	100	81-83	2	100		
156-59					78-80	4	99		
152-55	1	100	1	100	75-77	6	96		
148-51	3	99	1	99	72-74	10	92		
144-47	5	97	2	99	69-71	12	86	2	100
140-43	12	94	5	98	66-68	20	78	2	99
136-39	13	87	7	95	63-65	26	66	7	98
132-35	21	78	17	92	60-62	22	49	8	94
128-31	21	65	19	83	57-59	23	35	17	90
124-27	24	52	17	74	54-56	8	20	27	82
120-23	18	36	21	65	51-53	12	15	31	64
116-19	13	25	12	54	48-50	7	8	36	53
112-15	9	17	28	48	45-47	2	3	17	35
108-11	5	11	17	34	42-44	3	2	25	26
104-07	3	8	20	26	39-41			10	14
100-03	5	6	10	16	36-38			15	9
96-99			10	11	33-35			2	2
92-95	2	3	2	6	30-32				
88-91	2	1	6	5	27-29			1	1
84-87			3	2					
80-83									
76-79			1	1					
Q_3	135		128			67		55	
Md	127		117			63		50	
Q_1	120		107			57		44	
Mean	126.3		117.2			62.5		49.8	
σ	12.21		14.40			8.16		7.63	

difference might be explained by differences in the abilities of the groups of students in the two schools were it not for the fact that on the remaining sections of the examination, Navigation, and Ordnance and Gunnery, such differences did not appear.

IMPROVEMENT IN GRADING AND MARKING. Two types of evidence indicate the improvement in grading and marking which resulted from the introduction of final achievement examinations in the Tactical Radar School: (1) the interrelationship between various grades before and after the introduction of the tests, and (2) the increased efficiency of prediction of school success when the final achievement examination grades are used as a criterion or as a factor in the criterion.

The final course grades for the first three classes completing the course of the Tactical Radar School consisted of the weighted average of the "theory" and "practical" grades assigned each student. The "theory" grade, contributing one-third to the final average grade, consisted of the arithmetical average of all grades received in weekly quizzes. The "practical" grade was based upon ratings of the student's performance in the Combat Information Center mock-up. Original ratings were made on a scale ranging from 1, high,

TABLE 4-XVI. Intercorrelations among grades comprising the Final School Grade for the fifth class at the Tactical Radar School

	Achievement Examination Grade	Theory Grade	Practical Grade
Theory Grade	.59		
Practical Grade	.11	.13	
Final School Grade [1]	.80	.89	.31

[1] This grade is the arithmetic average of the other three; therefore the correlation coefficients in this row are spuriously high.

to 5, low, on the traits of (1) leadership, (2) teamwork, (3) judgment, (4) mental agility, (5) surface plotting, (6) air plotting, and (7) speech. After combination, these ratings were converted to grades expressed in terms of the Navy "4.0" system. These "practical" grades made up the other two-thirds of the final average grade. For the fourth and fifth classes the results of the final achievement examination were introduced in the final grades. The students' scores on the CIC Final Achievement Examination were converted into Navy "4.0" grades and these were averaged together with the "theory" and "practical" grades which were determined in the same manner as with the previous classes, each of the three grades being given a nominal weight of one.

The interrelationships of the grades used in computing the final school grades for the fifth class are shown in Table 4-XVI. Particularly striking in Table 4-XVI are the low correlation coefficients between the "practical" grade and both the "theory" grade and the achievement examination results.

It should be remembered that the "practical" grade consists of the ratings which were made of the student-officer's performance in the Combat Information Center mock-up. While one would not expect a perfect correlation between such a measure and performance in class work, it does not seem reasonable that the relationship should be as low as that shown in this table. If knowledge about a subject contributes to performance, one would expect the relationship to be higher, since the "practical" grade and the component ratings purport to evaluate the officer's performance in a Combat Information Center, and the development of competence in such performance was the general objective of the entire training program of this school.

The final school grades for classes seven, eight, and nine were somewhat differently derived. The "theory" grade was based upon the course grades for the first four weeks. The "practical" grade

TABLE 5-XVI. Intercorrelations among the grades comprising the Final School Grade for the ninth class at the Tactical Radar School. (N = 117)

	Achievement Examination Grade	First Month Grade	Second Month Grade
First Month Grade (Theory)	.78		
Second Month Grade (Practical)	.66	.69	
Final School Grade (Average) [1]	.91	.93	.84

[1] This grade is the arithmetic average of the other three; therefore the correlation coefficients in this row are spuriously high.

included the grades made during the final four weeks of the course and the grade obtained from performance in a comprehensive practical examination. The grade on the CIC Final Achievement Examination was the third factor in the arithmetical average which constituted the final school grade for these classes. The intercorrelations of these grades for the ninth class are shown in Table 5-XVI.

The data in Table 5-XVI, when compared with those in Table 4-XVI, indicate that the elimination of the rating of performance traits in the Combat Information Center mock-up, which had originally accounted for two-thirds of the final grades, resulted in a substantial increase in the intercorrelations of the grades. The correlation coefficients of .66 between the achievement examination scores and second month grades and of .69 between the grades for the first and second months correspond more nearly to what one

would expect correlations between different measures of success to be. Apparently the introduction of the final achievement examination and the comprehensive practical examination brought about a better definition of the criterion of success to be employed at the Tactical Radar School, as evidenced by these higher interrelationships among the grades making up the final grade assigned to each student.

IMPROVEMENT IN THE EFFICIENCY OF PREDICTION. Increased efficiency in prediction of school success, an important element in the reduction of attrition, is indicated by the data in Table 6-XVI which compares the correlation coefficients between scores on parts of the Officer Classification Test and the grades of students before and after the introduction of the achievement examinations.

For class three, the only relationship which attains the five per cent level of significance is that between Part I—Verbal Section of

TABLE 6-XVI. Correlation coefficients between scores on the Officer Classification Test and criterion measures at the Tactical Radar School, for classes on which the bases of grading differed

	Class 3, N = 178			Classes 7, 8, 9, N = 83	
Officer Classification Test Part	Theory Grade	Practical Grade	Final Grade	Achievement Examination Grade	Final Grade
I Verbal	.16	.13	.18	.31	.32
II Mechanical	—.06	.07	.04	.30	.19
III Mathematical	.07	—.04	.01	.44	.49
IV Spatial	—.09	.01	—.02	.13	.08

the Officer Classification Test and the final school grade. For classes seven, eight, and nine, the correlation coefficients of both the Verbal Section and Mathematical Section of the Officer Classification Test with both final school grade and achievement examination are well above the one per cent level of significance. It is also to be noted that the correlations with the achievement examination are generally somewhat higher than those with the final school grade. That this improvement in prediction is not merely a statistical artifact is indicated by the changed pattern of relationships. The most striking fact is the high relationship between the Mathematical Section of the Officer Classification Test and both the achievement examination score and the final school grade. Evidently the introduction of the final achievement examination resulted in a greater emphasis being placed on the technical and navigational phases of the curriculum. Whether the final grade which resulted from the use of the new

grading factors, including the achievement examination and eliminating the ratings, constituted a more valid measure of the success of student officers is of course a different question. It can be said, however, that the new criterion was predictable by means of the Officer Classification Test whereas the old one was not. It is important to note that in the judgment of the school staff the new criterion was superior to the old one.

The Future of Achievement Tests for Officer Schools

Much of the material prepared for use in the wartime officer training schools will prove useful in the officer training program of the post-war Navy. Present plans indicate that an examination program for the Naval Reserve Officer Training Corps units will be developed when the standard curriculum for the NROTC is adopted and put into operation. A series of tests is now in preparation for use in the graduate general line school which is to provide additional training for the reserve and temporary officers who will be appointed to commissions in the regular Navy. It is also expected that the examination materials, modified and augmented for the purpose, will be made a part of the examination stockpile from which examinations for promotion are prepared.

In a sense, the examinations developed for the officer training programs during the war can be considered the fragmentary application of principles and techniques of achievement testing which have been developed in American colleges and universities over the past twenty years. Even though the examination program was developed too late in the war to demonstrate the full potential utility of measurement in the evaluation and control of the training programs for officers, enough of a start was made to indicate the need for a continuing program of research and measurement, both as a service to the officer training establishments of the peacetime Navy and as an element in maintaining the readiness of the Navy to conduct expanded and intensified training adapted to the specialized needs which any future emergency may raise and define.

Gunfire can be most effectively brought "on target" when observers can report whether successive salvos are under or over, left or right. Similarly training programs can be more effectively directed toward the accomplishment of their objectives when means of measuring the outcomes of instruction are provided and applied. It is expected that the post-war program of training research will undertake the systematic development of measuring instruments for the evaluation of training.

CHAPTER XVII

THE MEASUREMENT OF ACHIEVEMENT IN THE RADIO TECHNICIAN TRAINING PROGRAM

THE Navy programs for training enlisted personnel to operate and maintain electronic equipment have paralleled the phenomenal development of various types of radio, radar, sonar, fire control, and related gear. Before the war, when the uses of radio and related electronics gear were considerably less developed, the majority of naval radiomen were assigned to routine communications duty. Upkeep of the electronic equipment was the responsibility of the chief and first class petty officers (radiomen) who were graduates of the single Radio Materiel School or who had shown a special aptitude for maintenance work.

As early as 1928, however, communications personnel in the fleet had become increasingly aware of the necessity for training men specifically to maintain and repair the operating equipment. By 1939, it was clearly apparent, not only in the fleet but also in the cognizant bureaus, that thousands of men (later designated as radio technicians and more recently as electronic technician's mates) were needed to service the equipment which was multiplying in variety and becoming increasingly complex. Between 1939 and 1941, plans for augmenting the training programs were underway and a building program was initiated. In December 1941, classes averaging 75 men were entering the training program every two weeks. The tremendous demand for personnel to be trained to meet the maintenance needs of a rapidly expanding fleet, coming at the time when all the personnel programs of the Army, Navy, and industry were expanding, gave rise to a number of problems.

THE PROBLEM OF SELECTION. Before and during the early months of the war, only men who had some previous background in radio and electrical work were accepted for training as maintenance personnel. But as a result of war pressures, personnel input quotas to the program were increased. One immediate result was the exhaustion of the supply of experienced men and the subsequent necessity of selecting for training from the general recruit population large numbers of men whose acquaintance with radio usually stopped with the ability to turn on and off a home receiver and to tune in desired programs. As selection standards were lowered, many of the men selected were unable to maintain the pace of the intensive and accelerated program. The rate of attrition became very high. Im-

proved selection instruments and procedures were urgently needed to identify among the available men those who, when trained, would be most serviceable to the Navy. This led to the development of the Radio Technician Selection Test which has been described in Chapter VIII.

THE PROBLEM OF TRAINING. Furthermore, it was found that many of the capable men did not have the background in mathematics, engineering, and electricity required for successful completion of the course of instruction in Radio Materiel school. To supply this background, two types of preparatory schools were established. The first of these was known as the Elementary Electricity and Radio Materiel (EE and RM) schools. As the program expanded and as the induction of inexperienced personnel increased, schools of this level were also found to be inadequate to supply all the necessary elementary training. Therefore, the one-month Pre-Radio-Materiel schools were established on the preliminary level. The training program as thus organized, consisted of three levels of schools, Pre-Radio-Materiel, EE and RM, and Radio Materiel, covering a total period of eleven months.

During the course of instruction in Pre-Radio-Materiel schools, an attempt was made to give men of very different backgrounds a basic knowledge of certain elementary concepts of mathematics and electricity believed essential for successful performance in EE and RM schools. In large measure this program consisted of intensive review of high school mathematics and physics.

Graduates of the Pre-Radio-Materiel schools were sent to EE and RM schools for three months of classroom and laboratory training in mathematics, electricity, rotating machinery, and radio theory. This curriculum represented new learning for almost all trainees.

Graduates of the EE and RM schools were then considered prepared to undertake the study of Navy electronic equipment as taught in the seven-month advanced course at Radio Materiel schools.

With a great increase in the length and volume of radio technician training, standardization of the program became a problem of major importance.

The Improvement of Training

In December 1943, plans were made to double the input of men to the radio materiel training program; and subsequently, as needed, new schools were opened. A large number of these were on the second level (EE and RM), and of these more than half were located in civilian colleges and engineering or technical schools under contract with the Navy.

Navy equipment was continuously subjected to objective and closely controlled tests to establish the superiority of one piece of gear over another; but in the field of personnel, little evaluation of training of radio technicians had been made beyond the pragmatic observation that, in general, the type of training being received in existing Navy schools appeared to be adequate. Therefore, with the expansion of training, the new contract schools were given an outline (often meager) of what the existing program had been in the Navy schools. From this the school faculties were expected to formulate a curriculum covering the fundamental concepts of electricity and radio.

It is readily understandable that the variety of programs thus developed in these contract schools resulted in a non-uniform type of graduate, a constant source of trouble in subsequent training. Inevitably, the first month of instruction in the advanced schools had to be spent in review and even first teaching of certain fundamentals in order to make sure that each man had the necessary preparation to undertake the advanced curriculum.

Because of this lack of uniformity in school graduates, the Electronics Section and the Test and Research Section of the Bureau of Naval Personnel early in 1944 undertook jointly to set up a comprehensive standardization program for the Pre-Radio-Materiel and EE and RM schools. Previous curriculum studies were reviewed, practices of the existing schools were examined, and the needs of the advanced schools were analyzed. The curricula and laboratory outlines for the Pre-Radio-Materiel and EE and RM schools were prepared and placed in operation in March 1944.

There had been a long tradition of much mathematics in the radio materiel training program. Most of this content was concentrated in the EE and RM schools and was presumably necessary theoretical background required for success in the advanced schools. But as the training program became more and more concentrated, the necessity of much of the mathematics taught was frequently questioned. Considerable opposition, based on the opinions of instructors and on reports from the fleet, developed to the practice of including so much mathematical material in the curriculum. One special study was carried out to obtain a more objective picture of the nature and extent of mathematics required by the advanced schools. To provide specific information, a check list of mathematical concepts was prepared and submitted to the instructional staffs of the schools. From the combined reactions of the instructors it was possible to recommend that certain large areas of mathematics in the EE and RM curriculum could be omitted entirely, while others could well receive less emphasis. The ensuing changes caused no

perceptible lessening of trainee quality. As a matter of fact, there was a subsequent marked trend toward increased practicability of curriculum even in the advanced schools.

THE DEVELOPMENT OF ACHIEVEMENT EXAMINATIONS. Two series of standardized final achievement examinations were developed to implement the new Pre-Radio-Materiel and EE and RM curricula. From the beginning, these examinations were regarded as an integral part of the training program. It was felt that no matter how detailed the curricular outlines and syllabi might be, the Pre-Radio-Materiel and the EE and RM schools could also get direct leads as to the type of preparatory training the advanced schools deemed necessary from the examination hurdles set for the graduates of the two preparatory schools. To this end each series of examinations was carefully designed to sample in a comprehensive manner the entire content of the respective curricula.

Specifically, the aims of achievement examinations in these schools may be stated as follows:

1. To direct instruction toward curricular objectives and to serve as a basis for necessary changes in the courses of study.
2. To check on adequacy and quality of instruction.
3. To supply an objective and comparable measure of achievement of individual trainees, classes, and schools.
4. To provide a final test mark as a basis for declaring inaptitude, for holding over for further instruction, or for graduating trainees to the next phase of the program.

Pre-Radio-Materiel Achievement Examinations. Six forms of the Pre-Radio-Materiel Achievement Examinations were constructed. These tests consist of four-response items. Forms 3-6 are made up of 100 items each, distributed as follows: Mathematics, 35 items; Electricity, 50 items; and Mechanical Practice, 15 items. Testing time is two and one-half hours. The dispersion of scores on these tests is uniformly narrow and the means are uniformly high, reflecting the fact that the Pre-Radio-Materiel curriculum consists in large measure of intensive review and that the examinations have been devised to measure minimum essentials.

Insofar as possible, items were developed at the level of application of principles to practical problems, but since much of the learning in Pre-Radio-Materiel schools is at the vocabulary level, such functional items have not always been practicable. Some experimentation was done with items of this type:

Directions:

In each of the two following exercises, the five steps to be performed in an operation are arranged in mixed order. Mark the

numbered answer spaces on the answer sheet to indicate the order number from 1 to 5 in which each step should be performed. In the sample below, b is step 1, d is step 2, etc.

SAMPLE

Steps in a soldering operation

a. Tin and clean the soldering iron
b. Complete the work to be soldered
c. Begin heating the soldering iron
d. Select an iron appropriate for the job
e. Apply heated solder to the joint

Answer Form

Steps



71-75. Steps in a drilling operation (Use spaces 71-75 on Answer Sheet):

71. Center drill in chuck and tighten
72. Determine size and location of hole to be made
73. Obtain drill of proper size
74. Center-punch proposed hole
75. Fasten on support material to be drilled

76-80. Steps in soldering wire to a smooth surface (Use spaces 76-80 on Answer Sheet):

76. Turn off heating current
77. Hold wire on plate, apply heat, and solder
78. Push back insulation and clean surface to be soldered
79. Remove iron and hold connection rigid
80. Plug iron into receptacle to allow to heat

It was hoped that understanding of correct procedures could thus be measured. Subsequent experience showed, however, that such items were subject to many ambiguities, since in many operations there is no single proper sequence. When such items were constructed free from ambiguity, it was generally found that they had lost their discriminatory power. Although these items failed in their intended purpose, they did call the attention of instructors to the fallacious instructional emphasis upon a single acceptable order of procedure where, in a given situation, several different sequences might be equally correct.

EE and RM Final Achievement Examinations. The six forms of the final achievement examination for the EE and RM schools were constructed to sample trainee learning in the following areas:

Part I. Applied Mathematics, Fundamentals of Electricity, Fundamentals of Radio.

Part II. Communications Circuits, Power Supplies, Electrical Machinery, Fundamentals of Radio.

Part I (40 items in Form 1 Revised, 50 items in Forms 2-6 inclusive) yields a single raw score, Part II (80 items) yields four raw scores to facilitate diagnostic evaluation of the student's work.

Since the emphasis in the curriculum at this level is on the development of understanding of fundamentals, it was felt that the speed factor should be minimized in the determination of examination scores. The time limit of three hours for each part was established through a series of experimental administrations to determine the limit within which nearly all men could finish the examinations.

All test items are of the five-response multiple-choice type. An effort has been made to sample the trainee's ability to apply theoretical principles to the solution of practical problems. For example, the curriculum for the first and second months of the EE and RM school stresses the mathematics necessary for understanding the various characteristics and phenomena of alternating current. The test items in the examination, while essentially electrical in content, have been devised so as to sample practically all of the mathematical skills which the EE and RM graduate should have mastered as the basis for advanced school studies.

Throughout the examinations, the emphasis has been placed on problems which demand the ability to reason in terms of the facts learned. Typical diagrams and schematics, such as will actually be encountered when servicing gear aboard ship, are employed. Test items based on such schematics include the location of typical causes of faulty operation, the prediction of faulty operation which might result from various typical equipment failures, and the like. Such problems are believed to approximate fairly closely both the actual kinds of situations which might be encountered aboard ship, and the types of reasoning demanded for their solution.

Statistical treatment of each form of the EE and RM Final Achievement Examination has included determination of:

1. estimated item-test coefficients of correlation,
2. difficulty values,
3. Kuder-Richardson reliability coefficients,
4. part-whole and interpart correlation coefficients,
5. validity coefficients,
6. norms.

Table 1-XVII shows the average item-test correlation coefficients, difficulty values, and reliability coefficients for the six forms. While the reliability coefficients are not as high as those reported in Chapter VIII for the Radio Technician Selection Test, it is believed that they are satisfactory, particularly in view of (1) the tendency of the Kuder-Richardson formula to underestimate, and (2) the fact that the tests are given under no time-limit conditions. Items in each

form have been selected to provide a good range of difficulty values. At the outset are placed items which are solved by 90 to 100 per cent of the men. As a rule, approximately two-thirds of the items on each form are answered correctly by 50 per cent or more of the men. Only a few items are missed by as many as 75 per cent of the trainees. Despite the generous time limits for administration, all forms yield essentially symmetrical bell-shaped distributions of scores.

To determine the validity of the part-score breakdown of Part II of these examinations, interpart and part-whole correlation coefficients were computed for each form. Typical of the results are the data for Form 6 as shown in Table 2-XVII. It will be observed that

TABLE 1-XVII. Reliability coefficients, average item-test correlation coefficients, and average difficulty values for EE and RM Final Achievement Examination

Form Number	Number of Items	Reliability Data				Item Analysis Data	
		N	r(K-R)[1]	M	σ	Average Item-Test Correlation Coefficient[2]	Index of Average Difficulty Value[3]
1 Revised	120	200	.85	95.62	11.27	.36	54.3
2	130	200	.84	100.83	11.49	.34	76.5
3	130	200	.86	106.98	11.34	.38	81.5
4	130	200	.87	98.65	13.30	.42	73.8
5	130	200	.87	99.83	12.99	.39	76.4
6	130	464	.87	96.38	13.67	.40	77.7

[1] Computed by Kuder-Richardson Formula No. 21.

[2] Item-test correlation coefficients estimated from upper and lower 27% of the test score distribution. Flanagan, J. C., "General Considerations in the Selection of Test Items, etc." *J. Educ. Psych.*, 1939, 30, 674-680.

[3] Index of difficulty is defined as per cent of persons selecting the correct response to the item.

the interpart coefficients of correlation are consistently low when considered in relation to the part-whole coefficients. Thus the use of part scores appears to be warranted.

As previously indicated, the EE and RM schools are preparatory for the work of the Radio Materiel schools. The EE and RM Final Achievement Examination has been established as the criterion for determining pass-fail marks in EE and RM school. Therefore, a measure of the validity of these examinations is the extent to which they predict achievement in the third-level Radio Materiel schools. Table 3-XVII presents correlation coefficients between scores on four forms of the EE and RM Final Achievement Examination and marks in the first month of Radio Materiel school. These coefficients

are within the limits usually considered satisfactory for educational prediction.

Form 1 Revised, the only form available, was administered to classes graduating during the period June 23 to September 27, 1944. Because the scores increased steadily it was suggested that this form had probably been compromised by student gossip and by direct

TABLE 2-XVII. Part-whole and interpart correlation coefficients for EE and RM Final Achievement Examination, Form 6 (N = 200)

	Total	Part I	Part II	IIa	IIb	IIc	IId
Part I	.86		.50				
Part II	.88						
IIa	.64				.32	.31	.40
IIb	.63					.32	.43
IIc	.52						.38
IId	.73						
Possible Maximum Score	130	50	80	25	20	10	25
Mean	106.96	39.37	67.50	20.88	17.53	8.90	20.32
σ	9.14	4.73	5.71	2.25	1.80	1.22	2.42

TABLE 3-XVII. Correlation coefficients between scores on EE and RM Final Achievement Examination and marks at end of first month of Radio Material School

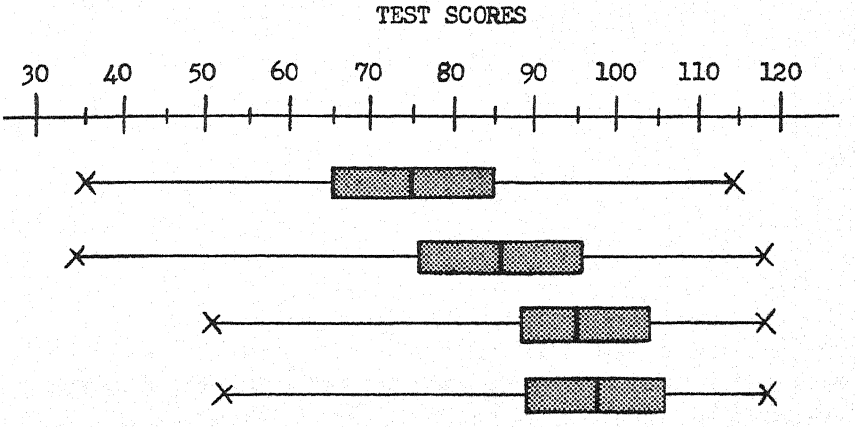
	EE and RM Final Achievement Examination							
	Form 1 Revised		Form 2		Form 3		Form 4	
School	r	N	r	N	r	N	r	N
W	.59	98	.55	102	.64	90	.66	88
X	.38	293	.52	273	.53	293	.65	317
Y	.49	203	.54	201	.54	175	.60	203
Z	.52	177	.59	184	.69	170	.70	198
Average r[1]	.47		.55		.59		.65	

[1] The average correlation coefficient is determined from the weighted z values. Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill Book Company, Inc., 1940, p. 155.

instruction for the examination. Therefore as soon as they became available, Forms 2 and 3 were administered to the classes of October 12 and 26 respectively. On November 8, Form 1 Revised was again administered; the scores on this administration still indicated positive improvement. The steady upward trend of scores on Form 1 Revised, which continued after additional forms intervened, has

been interpreted to mean that instruction has been directed toward curriculum objectives and that loss of security of examination contents has been negligible (Figure 1-xvii).

It has been possible to trace directly the effects of these examinations upon the instructional program. With the passage of time, item difficulty values tended to decrease in critical areas of subject matter. Numerous reports were received from school officials indicating that the examinations provided them with guide posts which enabled them to interpret the curriculum more fully. The sampling of instructional material represented in the examinations is so complete that if schools do teach for the examinations, they will of



Date of Administration. July 6th, August 2nd, September 13th, November 8th.

1. Heavy bar = score area covered by middle 50% of men.
2. Heavy line within bar = school average.
3. X–X = highest and lowest score.

Figure 1-xvii. Group gains in scores on EE and RM Final Achievement Examinations, Form 1 Revised (Forms 2 and 3 of this examination were administered to classes graduating October 12th and 26th, 1944).

necessity teach the desired materials with the intended functional-practical emphasis.

Perhaps the most convincing evidence of the standardization and improvement accomplished through the construction of standardized curricula and examinations came from the reports of the advanced schools. They were no longer able to identify men by the EE and RM schools attended because their skills were so uniform that the former individual school training peculiarities had disappeared; furthermore, achievement of the trainees improved sufficiently to eliminate the need for extensive review in the first month of advanced school and to permit the immediate undertaking of advanced instruction.

EE and RM Performance Tests. A project for developing comprehensive performance tests for the measurement of laboratory skills in the EE and RM schools was initiated early in 1945. The procedure followed in the formulation of these tests is similar to that described in Chapter XV. Representatives of the Test and Research Section visited each of the schools and assisted the local staffs in determining the appropriate tasks and skills to be measured and the possible techniques for evaluating performance. As a consequence, each of the schools undertook the development of experimental forms of laboratory performance test items which they incorporated as an integral part of their laboratory program.

In July a conference was called at the Bureau of Naval Personnel for coordinating this program. Technical officers of each of the EE and RM schools met with representatives of the Test and Research Section and other cognizant sections of the Bureau of Naval Personnel. The whole philosophy of performance testing was reviewed and discussed. Sample items developed by individual technical officers were analyzed for relative strengths and weaknesses. Following a survey of the curriculum, decisions were reached regarding the number of performance tests to be developed and the areas of the curriculum to be covered by each. Responsibility for the development of performance tests covering specific areas of the curriculum was delegated to each of the technical officers present. These projects were well underway when the surrender of Japan and subsequent separation of many of the technical officers brought a temporary suspension to this program.

Rating Scales for Instructor Training

Throughout the war the radio technician training program had its own instructor training school. Selected graduates from the advanced schools were given an eight-weeks course in teaching methods and materials, and upon the completion of this course were assigned to instructional duties in the EE and RM schools. Two rating scales were developed for use in this program.

THE TEACHER TRAINEE RATING SCALE. This scale was developed for the use of the supervising teachers in estimating the efficiency of the teacher trainees during their practice teaching period. It consists of twenty items considered by supervisors in the instructor training school to be essential to good teaching. Examples are: knowledge of material; organization of material; reaction of class to lecture; ability to handle questions; maintenance of schedule; use and choice of visual aids; poise; discipline of and relation to students. Each item is followed by five descriptive phrases, one of which is to be

checked as being most descriptive of the trainee being rated. A portion of the scale is shown below:

Directions: For each item check that statement which, in your opinion, most nearly fits the above man at the time of this report.

Knowledge of material:

__a. prepared; knowledge adequate
__b. poorly prepared; sometimes stumbles over material
__c. has knowledge of complex material; sometimes speaks above understanding level of students
__d. knows complex material; presents work clearly; easy to understand
__e. unprepared; knowledge inadequate

Organization of material:

__a. fairly well organized; some novel planning
__b. organized for most effective presentation
__c. incompletely adequate; slipshod
__d. organization adequate; conventional
__e. disorganized; confused

Reaction of class to lecture:

__a. interest fluctuates
__b. bored, restive, disinterested
__c. shows lack of interest
__d. engrossed, attentive, interested
__e. shows unusual absorption

Ability to handle questions:

__a. comprehends rapidly; gives answers; proceeds on course
__b. not disturbed; answers lengthy and indirect
__c. stimulates questions; answers effectively; does not leave main course
__d. disturbed by questions; leaves main course
__e. disturbed at first; regains poise and proceeds on course

During the course each trainee was observed and rated at least once a week, after which the supervisor discussed his ratings with him. Then on a self-keying profile the trainee plotted the ratings he was given; he was thus enabled to evaluate his progress from week to week.

The final assessment of the trainee's potential quality as an instructor was based on his status at the end of the course, on the assumption that his worth to the program was best represented by his final status rather than by an average which might reflect his

beginning inexperience. Although all of the instructor trainees showed improvement on this scale during their eight-weeks training program, the previously experienced groups attained maximum level earlier and on the whole were superior to the previously inexperienced groups. This is of course a natural expectation, since previous experience in teaching probably reflects a higher degree of interest in, and aptitude for, teaching.

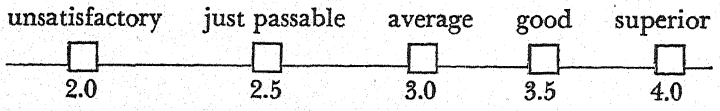
THE INSTRUCTOR RATING SCALE. The Instructor Rating Scale was constructed for use as a supervisory instrument in evaluating the quality of teachers in the EE and RM schools of the program. Ten aspects of teaching were selected, and a series of questions relating to each of these aspects was checked by the rater on a five-point scale. Sample items follow:

> *Directions:* Under each of the following ten headings are questions suggesting some types of skills which should be considered in evaluating teaching efficiency. Read the sample questions under each heading. Then place a check (√) in the box under each heading which indicates your judgment of the quality of work being done by the above-named instructor. In making your judgment on a particular item, disregard every other item. Do not attempt to mark each of the sub-questions individually. Your mark should represent your judgment of the man's overall performance under each heading.

1. *Lecture*

How effective is he in getting his ideas across to students?

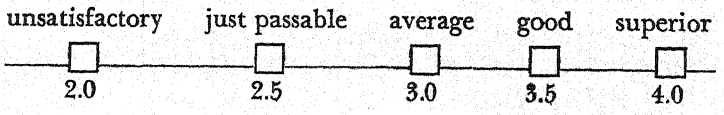
Does he make main points stand out? Are his lectures clear-cut and easy to follow? Does he define new terms when presenting them? Does he check student understanding during a lecture? Does he speak effectively?



2. *Laboratory*

How effective is he in directing laboratory work?

Does he keep students working efficiently throughout entire period? Is class production high? Does he have equipment ready for class use? Does he check equipment before use and keep it in good repair?



Summary

Coordination of effort has characterized the program for the selection and training of Navy radio technicians. The Test and Research Section has worked closely with Curriculum, Instructor Training, and Electronics Sections, with Training Aids Division of the Training Activity, and with the Classification Section of Enlisted Personnel Activity, in the development of selection and achievement tests, in the determination of selection procedures and standards, and in the development of curricular and training aids. This close cooperation of the various programs within the Bureau of Naval Personnel, working together with the staffs in the schools, has contributed greatly to the successful training of thousands of radio technicians for the Navy.

CHAPTER XVIII

ADVANCEMENT IN RATING EXAMINATIONS

As indicated in Chapter V, the standards for petty officer rating and advancement of enlisted personnel are established by the Bureau of Naval Personnel. All petty officers are expected to qualify in certain "petty officer qualities" and to possess the fundamental knowledge required of all men in the Navy. In addition, each petty officer is expected to meet the technical qualifications of his specialty. These technical qualifications are standard for each rate (pay grade) within the rating.

To illustrate the nature of the qualifications specified for a rating, there are listed below the specific standards which must be met before a man is rated radio technician third class. Before a man is advanced to radio technician second class he needs to be more expert both in practical matters and in examination subjects. A radio technician first class must meet still more technical standards, while to be rated chief radio technician, in addition to completing practical on the job training, a man is expected to have graduated from an advanced radio materiel school approved by the Bureau of Naval Personnel. The studies of a chief radio technician are so technical that it is unlikely that he would develop the necessary competence solely through shipboard or shore station training.

Qualifications for Radio Technician Third Class

A. Practical Factors

1. Primary Factors (Petty Officer Qualities)

(a) Leadership. Demonstrate the qualities of a good leader when in charge of a group of men. Demonstrate ability to train subordinates.

(b) Authority and Responsibility. Demonstrate ability to exert authority and maintain discipline under all conditions. If so assigned, demonstrate a knowledge of the duties and responsibilities of a shore patrolman or police petty officer.

(c) Security. Demonstrate knowledge of regulations regarding the security of classified matter.

(d) Drill. Demonstrate ability to take charge of a group and conduct infantry drill and physical exercises.

(e) Damage Control. Demonstrate knowledge of watertight integrity and material conditions. Demonstrate knowledge of the uses of all fire fighting equipment and rescue-breathing apparatus.

(f) First Aid. Demonstrate ability to administer first aid, including artificial respiration.

(g) Arithmetic. Demonstrate a fundamental knowledge of elementary arithmetic, including addition, subtraction, multiplication, division, simple and decimal fractions, and measurement.

(h) Sound-Powered Telephones. Demonstrate ability to man a telephone station, handle the equipment correctly, and follow standard talking procedure.

(i) Chemical Warfare. Demonstrate an elementary knowledge of chemical warfare and defensive measures.

(j) Surface Preservation. Demonstrate knowledge of painting regulations, paints, and painting equipment.

(k) Survival. Demonstrate knowledge of the proper method of abandoning ship, swimming in oil-covered water, and the use of life-raft supplies and equipment.

2. Technical Requirements

(a) Equipment Adjustments. Start, stop, regulate, and make necessary operating adjustments on the radio transmitting, radio receiving, radio direction finder (including HFDF), sonar, radar, frequency measuring, and vacuum tube testing equipment in own ship or station, and know the safety precautions involved. This includes the shifting of frequencies within the time limit and degree of accuracy set as standard within the fleet or force to which attached. When specifically assigned, perform the adjustments listed above on Loran and Countermeasures equipment.

(b) Circuit Diagrams. Demonstrate ability to draw and to interpret schematic circuit diagrams of simple electronic circuits.

(c) Identification of Components. Demonstrate ability to locate and identify component parts of an actual piece of electronic equipment by reference to the associated circuit diagram.

(d) Tools. Demonstrate ability to handle properly ordinary tools used in routine electronic service work.

(e) Simple Repairs. Demonstrate ability to make simple repairs to standard shipboard electronic equipment under qualified supervision.

(f) Remote Control Systems. Demonstrate a working knowledge of the electric and electronic remote control systems used with equipment in own ship or station.

(g) Power Supply Systems. Demonstrate a working knowledge of the power supply systems used with the electronic equipment in own ship or station.

(h) First Aid. Demonstrate a thorough knowledge of first aid with emphasis upon treatment for personnel suffering from electric shock and burns.

B. Examination Subjects

1. Technical

(a) Batteries. Have a working knowledge of the types, uses, care, and maintenance of batteries used in naval electronic equipment.

(b) Ohm's Law. Have a working knowledge of Ohm's law and be able to apply it in the solution of elementary problems.

(c) Electronic Equipment. Elementary knowledge of the purpose and use of electronic equipment used in own ship or station.

(d) Cathode Ray Oscilloscope. Elementary knowledge of the theory and operation of the cathode ray oscilloscope.

(e) Test Equipment. Working knowledge of standard naval test equipment used in servicing electronic and related equipment.

(f) Safety Precautions. Know the safety precautions to be observed with electronic equipment in own ship or station.

2. General Naval Subjects

(a) Uniform regulations and care of clothing.
(b) Discipline and offenses.
(c) Salutes and colors.
(d) Manual of arms.
(e) Drill.
(f) Handling of boats and oars.
(g) Knots and splices.
(h) Compass and relative bearings.
(i) Visual signaling.
(j) Swimming and life saving.
(k) General naval and shipboard organization.
(l) Advancements and awards.
(m) Enlistments, discharges, and pay accounts.
(n) Types, general characteristics, and nomenclature of naval vessels and aircraft.
(o) Daily routine in port and at sea.
(p) General drills aboard ship.
(q) Hygiene and first aid.
(r) Gas protective apparatus.
(s) Rust removal and painting.
(t) General safety precautions.

In-service training and the procedures followed in the advancement of men in the various enlisted ratings have been described in Chapter V. Rating and advancement are dependent upon (1) competence in the technical qualifications and general petty officer qualifications as described above, (2) length of service, and (3) the vacancies existing at the shore station or ship to which the man is

attached or in the naval service at large. The latter two requirements are largely dependent upon the exigencies of the service. In wartime, when personnel needs are great, the number of vacancies is large and the length of service required for promotion is reduced; in peacetime the number of vacancies is smaller and the period of service required for promotion is longer. But in peace and in war, emphasis is placed upon training men for positions of greater responsibility and on promoting them as they demonstrate their capabilities for positions which require greater technical competence and qualities of leadership.

Examining for Advancement in Rating

Responsibility for the in-service training or advancement of enlisted personnel rests with the supervising officers and with senior petty officers who supervise the study of appropriate training courses and give the necessary instruction in the practical factors of the rating. Responsibility for checking on the technical proficiency required is, in general, delegated to examining boards established at local activities which are authorized to develop and administer technical examinations to cover the examination subjects set forth in the qualifications for specific ratings.

This decentralized examining program has the advantage that members of the boards located at various naval stations are close to the actual scene of operation where they may have first hand information about the available equipment and operating procedures. The continual changes in apparatus and in duties in different locations have made this close contact with the job a matter of primary importance. On the other hand there are certain limitations to local development of examinations. During wartime even conscientious boards had too little time to develop effective examinations. Many activities lacked personnel who were trained in test construction, or who appreciated the importance of technically adequate examinations, or who possessed the skill necessary to construct such an examination. Then, too, the great number of local boards engaged in preparing tests led to an inefficient duplication of effort at a time when manpower needs were great. It was estimated at one time during the war that several thousand naval activities were constructing their own advancement tests, ranging from definitely poor to very good.

Scope of the Examinations

The Test and Research Section undertook to assist local boards in their examining functions by constructing test materials which

met accepted test standards and which introduced a greater degree of uniformity into the examining process. Series of examination questions designed to be used by the local examining boards were constructed, assembled into four books, and distributed to all ships and stations.

The first series of tests were issued in the summer of 1944 as *Examinations for Aviation Ratings.* Included were tests of from 60 to 90 items for each of four pay grades for fifteen aviation ratings. The use of the examinations was optional for local examining boards.

The examinations currently in use (1946) are included in three books of *Advancement Examinations.* Book I consists of materials for ratings in the Seaman Branch, Commissary Branch, and Special Branch; Book II includes materials for ratings in the Artificer Branch; and Book III, for ratings in the Aviation Branch. The complete list of examinations in the current series is given in Appendix B.

As stated in the introduction to each book of *Advancement Examinations,* the items contained therein are not designed to be used as tests, *per se,* but rather as source material from which selections may be made and to which other items may be added in the construction of examinations to be used at local stations. The books contain more questions than are needed for any one examination. The use of these items, furthermore, is not mandatory, since it is possible that in some instances only a few of the items included under any particular test topic (examination subject) may apply to the individual ship or station to which the candidate is attached. But the boards are advised that items should not be excluded on this basis alone if the requirements are such that the man is to be qualified for any and all duties of his rating. Local examining boards are instructed to prepare additional items when the qualifications require that the man give information concerning specific gear aboard his type of ship.

How the Examinations Were Constructed

Based on Qualifications for Advancement in Rating. Each test developed for the advancement examination books deals with the specific requirements established for a particular rating. These qualifications are compiled by the Billet Analysis Section after extensive observation of the duties performed by men of the various ratings in fleet and shore billets and as a result of consultation with the various technical bureaus concerned. When completed, these qualifications are published in the *Bureau of Naval Personnel*

Manual [1] for the guidance of ships and stations in the advancement in rating process. The qualifications for radio technician, third class, presented earlier, illustrate the type of description of qualifications. In building advancement examinations during the war, the Test and Research Section has been concerned with constructing items for examination subjects, although plans were laid to develop performance tests and check lists covering the practical factors also.

The qualifications as established by the Billet Analysis Section, therefore, furnished the framework for the advancement examinations. If the qualifications included seven different examination subjects similar to those described above for radio technician, third class, the final technical examination accordingly was developed to cover all of these. Examinations for the higher pay grades also included the examination subjects of the lower pay grades.

PROBLEMS IN THE CONSTRUCTION OF THE EXAMINATIONS. The first major problem encountered reflected the scope of the material to be covered. Following are some of the questions which arose in the item construction process: What information should be sampled? How applicable is a particular question to operating conditions? Is it inclusive enough? Is it too general? Does the question interpret accurately the complexity of the task or job tested? Does the test item refer to specific gear? Is the equipment properly identified?

A second problem inherent in the construction of examination items was the necessity for making each item as practical and concrete as possible. The use of items testing the application of general principles to specific instances likely to be encountered by the average examinee in his daily work was stressed. Items covering abstract, theoretical concepts, infrequently applied and less frequently understood by the examinee, were generally avoided. Items involving the memorization of names, dates, places, rules, or laws were kept at a minimum, but the practical applications of these factors to the job situation were considered of prime importance and were tested accordingly. There was, however, no hesitancy in using technical nomenclature or concepts in the phrasing of the questions, since these examinations were designed to evaluate the qualifications of technicians in specialized fields. These men were not only required to perform the technical duties of their billets but were to be entrusted with the responsibility of training other men. They had to be able to comprehend the terminology of their specialty. Items in their finally adopted form were designed to measure specific information about equipment and operations.

The third problem to be dealt with involved the selection of the

[1] "Qualifications for Advancement in Rating," *Bureau of Naval Personnel Manual*, Part D, Chapter 5, Section 2.

types of items to be used in the examinations. The field was narrowed to five principal categories: multiple-choice, matching, completion, true-false, and pictorial-matching items. The use of the pictorial-matching type was restricted because of the difficulties which would be encountered in the duplicating process at local stations. In general, more emphasis was placed upon the use of multiple-choice and matching items. Completion items were used sparingly because of the difficulties inherent in the formulation of a fair and accurate key for their scoring. In the final series, approximately 70 per cent of the examination items were of the multiple-choice type, 20 per cent were of the matching type, 2 per cent were of the completion type, and 7 per cent were of the true-false type, with the remaining 1 per cent being miscellaneous items which did not fall under any of the above categories. Table 1-XVIII shows the

TABLE 1-XVIII. Distribution of different types of items used in the Advancement Examinations, Books I, II, and III

	Type of Item										
	Multiple-Choice		Matching		Completion		True-False		Miscellaneous		
Book	No.	%	No.	%	No.	%	No.	%	No.	%	All Types
I	5,453	57.3	2,803	29.7	339	3.6	691	7.3	194	2.1	9,480
II	5,773	76.6	1,157	15.4	137	1.8	454	6.0	12	0.2	7,583
III	6,583	78.7	1,163	14.0	110	1.3	478	5.8	13	0.2	8,347
All Books	17,809	70.2	5,123	20.2	586	2.3	1,623	6.4	219	0.9	25,360

distribution of the types of items as incorporated in the examinations.

DETERMINING EXAMINATION CONTENT. Informational material used in the construction of the examinations was secured from various sources within the Navy. Requests for suggested items, for training materials, and for locally formulated tests covering pertinent examination subjects were addressed to service schools for enlisted personnel and to training commands. Supplementary information was secured through correspondence with a number of units of the fleet. The materials obtained in this manner provided valuable information from which appropriate items could be built. Frequent use was made of manuals prepared by the Training Courses Section in the Standards and Curriculum Division of the Training Activity. The test constructor also had access to material available in other activities in the Bureau of Naval Personnel and in the other bureaus of the Navy, as well as to achievement examinations

previously constructed by the Section. When any given area was not sufficiently covered by the material on hand, additional steps were taken to secure the appropriate information from other special sources.

Insofar as possible, each officer working on the examination project was assigned the responsibility of developing tests in those particular areas where his training and experience could be utilized to a maximum degree. A number of warrant officers and chief petty officers having extensive experience in the various subjects were also brought to the Bureau from appropriate field commands to help in the compilation of the questions.

MAINTAINING A TEST ITEM FILE. In order that the individual items of the various examinations might be kept available for ready reference, a filing system peculiar to the requirements of the advancement examination program has been compiled, each individual item (in the case of matching items, each group) was transferred to a single McBee Keysort Card, 5 by 8 inches in size. The following information was also posted on each card: name of examination, examination subject, edition of the examination, and item number. Appropriate spaces on the card were provided for information relative to the difficulty of the item, and its internal consistency or validity against an external criterion. These cards were routinely made for every item. Duplicate items in two different examinations were posted on the same card when they occurred within the same branch of the service, but if they appeared in different branches they were cross-referenced from branch to branch. The edges of the test item card are provided with a series of small round holes for coding some of this information. The top edge of the card is coded for branch and rate; the left edge, for the kind of test item (in this case, Enlisted Advancement Tests); the right edge for form or edition of the test; the bottom edge, for examination subject, and for pay grade within the rating.

It was realized that, although the cards would prove useful in new test construction even though they were coded only by branch and rate, additional service could be provided by using a code based on examination subjects. With such a sorting system, all items on optical instruments, for example, could be rapidly assembled from any of the rates which would be concerned. Several plans for coding the examination subjects were considered. The number of examination subjects in the qualifications for advancement in rating roughly approximated 2,000, but many of these subjects were so similar that it was difficult to determine which were different enough to warrant separate classifications. These examination subjects were accordingly arranged in a code system specifically for use in test construction.

Table 2-XVIII lists the main topics with their codes. Each test item could be coded for one or two of these two-digit code numbers.

TABLE 2-XVIII. Codes for examination subjects used in coding of items in the McBee Keysort System as set up by the Test and Research Section

Code Number	Code Group
10 to 19	Ships and Planes: Structure and Handling.
20 to 29	Ordnance.
30 to 39	Propulsion Machinery; Auxiliary Gear; Ship Layout.
40 to 49	Accessories (Miscellaneous).
50 to 59	Electricity; Electrical and Electronic Equipment.
60 to 69	Fabrication.
70 to 79	General Physical, Chemical, and Mathematical Principles (except Electricity, Magnetism, and Metallurgy).
80 to 89	Safety and Damage Control; Sanitation; Nutrition; Anatomy; Physiology.
90 to 99	Supplies; Reports; Records; Publications.
00 to 09	Communication and Organization.

The development of the code system and the original assignment of code numbers to the items has been done by the officers concerned with preparing advancement examinations. To facilitate the routine maintenance and use of the file, a coding manual designed for use with the McBee Keysort Card file was written to describe the posting, coding, and card utilization systems.

The cards are used frequently in the production of new tests. They have already proved successful in eliminating the rewriting of similar items by different test constructors, in controlling the number of similar items used in different editions of a test, in eliminating errors which occur from the endless series of typings and retypings often carried out, and in sorting out, for reference purposes, items already prepared on a topic.

EXPERIMENTAL TRY-OUT AND PRELIMINARY STANDARDIZATION OF THE FIRST SERIES. It was indicated earlier that the first work on advancement examinations was the preparation of short tests for aviation ratings. Experimental editions of these tests were administered at shore stations and service schools where enlisted personnel in these ratings were located. Ordinarily, a test for a given pay grade within a rating was administered to twenty or thirty individuals. Analysis of the responses made to individual questions in the tests revealed the poor items—those missed by many people, those passed by everyone, and those on which the wrong answer was chosen frequently by men who had consistently high scores on the rest of the test. As a result of these preliminary analyses, the individual tests were improved by the elimination or rewriting of ambiguous or im-

practical items. The revised series of tests were then published and distributed to the various aviation activities where each command was authorized to utilize the examinations as a basis for its own testing program, with the addition of locally formulated items if desired. It was also requested that completed examinations be returned to the Test and Research Section for further analysis and study.

CUTTING SCORES. Before the Examinations for Aviation Ratings were distributed for experimental administration, tentative cutting scores were established for each test. Most of the men taking the preliminary examinations received scores considerably higher than the cutting score, while only a few cases fell below the critical point. Many of the examinees had recently completed school training; hence it was assumed that their marks would be somewhat higher than those of a group who had been serving at sea for some time.

In actual operation, the cutting scores were established to identify those individuals who needed additional training in order to be considered qualified in the examination subjects. In determining the cutting score it was practically impossible to set a definite inflexible cut on any one examination and still permit a ship or station to maintain its complement, both because of the necessity of rating some men to take the places of those detached and because of the widespread distribution throughout the world of men taking the tests.

When Advancement Examinations, Books I, II, and III, were issued, no attempt was made to secure normative data or to set cutting scores after trial runs. Since these books were not designed as complete comprehensive tests but were considered to consist of pools of items from which the examining boards might select, and since neither the individual items nor the groups of items had been calibrated for difficulty, the resulting examinations as developed by the boards were characterized by variation both in emphasis and in level of complexity. Under these conditions it would not have been feasible to designate cutting scores for all activities. The previously established requirement of the Bureau of Naval Personnel, that the final technical examination must be passed with a mark of 2.5 (62.5 per cent), served to set a somewhat flexible standard for the local examining boards.

Evaluation

The development of advancement in rating examinations for each pay grade in 50 ratings represented one of the most ambitious testing programs undertaken by the Test and Research Section. Although the examination program falls short of ideal standards with respect to experimental try-out of tests and standardization and uniformity

of use, it did constitute a practical solution to a very difficult problem. One of the major obstacles to the development and use of the typical standardized achievement examinations has been the variety of equipment found aboard different types of ships and the varied nature of Navy jobs bearing the same name. This difficulty was met by delegating responsibility to the local board for building examinations and by requiring satisfactory performance on the types of equipment found at the local station or aboard a specific ship.

The work in the development of advancement examinations was not completed during the war. If the best qualified individuals are to be advanced in rating, the program will have to be greatly extended. Examinations to cover proficiency in practical factors as well as in examination subjects will have to be built. Wherever possible, only items of known difficulty and validity should be used. In peacetime, extensive experimental try-out of items should be feasible. By building examinations which provide a valid measure of the proficiency of personnel, the Navy will insure a highly qualified staff of petty officers capable of discharging effectively the varied duties required in their billets.

PART V

FOLLOW-UP STUDIES OF TRAINING AND CLASSIFICATION TECHNIQUES



CHAPTER XIX

PROBLEMS IN ESTABLISHING CRITERION MEASURES

In personnel research no subject is of greater importance than the nature of the criterion to be used in evaluating performance. The criterion provides a means of isolating variables associated with individual differences in success, of separating efficient test items, subtests, and tests from those less efficient, of determining effective weights to be used in combining tests, and of evaluating existing training and classification procedures. In short, the criterion determines to a large extent what the findings and recommendations of a research study will be, and hence it may influence greatly administrative procedures involving selection, classification, training, assignment, and promotion of personnel. Use of an unacceptable criterion in a research study may result in policies or procedures which will be injurious to the development of sound personnel practices. In view of this importance to naval personnel research generally, and in order to introduce the reader to some of the problems encountered, it is considered desirable to indicate the essential characteristics of an acceptable criterion measure, to outline some of the problems which are involved in establishing such measures, and to present a brief critical analysis of the criteria that have been used in Navy research studies.

Definitions. The term criterion is applied to a performance measure which is used as a standard in evaluating other measures. It is also a means of describing the performance of individuals on a success continuum. The term "measure" is applied to the result of a process by means of which symbols are assigned to an individual's performance to designate (1) the presence or absence of a discriminable characteristic, or (2) a value representing a rank order position on some assumed variable. Depending upon the definition of the success continuum, a criterion used in Navy studies may include one or more of such variables as an individual's final school grades, his scores on proficiency tests, his position in the enlisted pay grade scale, an appraisal of his ability to direct the work of others, a rating of his skill shown in performing his job, the quality and/or quantity of work produced, his expressed job-satisfaction, and the judgment of supervisors or peers as to his overall success. The measure finally adopted may be composed of ratings, of scores or responses from controlled test situations, or of both of these.

Basic Assumptions. The major assumptions to be made in con-

sidering the development of a criterion are (1) that a scale or continuum of success does exist for Navy billets, (2) that individual differences in success can be discriminated, and (3) that numerical values or categories can be assigned which for practical purposes will stand for the set of rank orders representing the observed individual differences in proficiency.

The first of these assumptions, that of a continuum of success, is often made, however, without an understanding of the multiplicity of factors contributing to the success variance of the population. Although it may be convenient to think of success as a unitary variate, it is nearly always multidimensional in character. Seldom, therefore, will a single variable provide the required degree of discrimination between individuals on the total success scale. A single composite value representing overall success as a radioman, for example, would involve consideration of such skills as receiving and sending code, operating equipment, using communication procedures, standing watch, supervising a group of apprentices, and working as a member of a team.

The complexity of a particular continuum of success is directly influenced by the objective for which the measure is to be obtained. For example, if a measure of overall success on the job is required, several dimensions may be required, as was indicated above in the radioman example. If, however, the definition of success is one of technical competence and skill in a restricted area, as proficiency in receiving code in the case of radiomen, the complexity of the criterion will be low and may be represented by a single variable. In general, the composite scale of success and the operations necessary for the measurement of each of the possible dimensions (in the sense of establishing acceptable categories of performance characteristics or of rank orders of success) are imperfectly defined. Hence, extensive study of the areas of performance to be included is usually necessary before the criterion of a personnel research project can be developed. The definition of the scale of success and of the operations required in its application pose a series of difficult problems for the research worker.

Characteristics of an Acceptable Criterion

A criterion measure is a means of evaluating the proficiency of individuals on a success continuum. As a performance measure, it should satisfy the standards which are ordinarily applied to psychological measuring instruments or techniques. An acceptable criterion must possess three important characteristics, two of which can be further subdivided. First, it must be relevant to job performance

and include all significant aspects of that performance; that is, it must possess the qualities of *pertinence* and *comprehensiveness.* Second, it must consistently discriminate between different levels of performance, thus fulfilling the requirements of *discrimination* and *reliability.* Third, it must produce *comparable* results from individual to individual, from group to group, and from situation to situation. Each of these characteristics will be discussed more fully below.

PERTINENCE AND COMPREHENSIVENESS. The first important requirement of an acceptable criterion is that it must be pertinent to the performance being studied. If one were interested in predicting the technical competence of a radioman, it would be a mistake to evaluate him on such factors as how hard he works, how well he gets along with others, or his military appearance. All too frequently, research workers accept whatever criterion is available and assume that it is relevant to the type of performance which they are interested in predicting. The exclusion from the criterion of all extraneous variables not strictly bearing on job performance should be one of the first steps in designing a research study which aims to predict success on the job.

A concept closely related to pertinence is that of comprehensiveness. Although an acceptable criterion must not include extraneous or irrelevant factors, it must include all of the important components of the areas of performance delimited by the definition of success. A measure may be pertinent; but if it deals with only one of a series of variables which contribute to the composite success variance of the performance area under scrutiny, it cannot, by itself, be considered a sufficient criterion. For example, it would be quite inappropriate to accept code receiving as the sole criterion of the technical competence of a radioman. Code receiving is relevant to the criterion but it is not the only factor which determines the technical competence of a radioman.

CRITERION VALIDITY. In test construction, the characteristics here called pertinence and comprehensiveness are usually subsumed under the concept of validity. A test is valid to the extent to which it measures that which it purports to measure. In many test situations, the validity of the test can be defined as the correlation between the test and another measure which has been accepted as a standard. Here validity can be considered as a statistical concept. But in the development of an original criterion of success, no such standard exists. If there were such a standard, one would adopt it as the criterion rather than spend his time developing a new one, unless the existing standard were too inconvenient for practical work or too gross to yield satisfactory discrimination. In the development of an

original criterion, as in the development of a test in an area for which no standard exists, validity is a logical rather than a statistical concept.

The two logical divisions of the *statistical* concept of validity are pertinence and comprehensiveness; the magnitude of a validity coefficient is a function of (1) the extent to which the thing measured by the experimental test is relevant to what is measured by the standard (pertinence), and (2) the extent to which the experimental test measures or includes that which is measured by the standard (comprehensiveness). In the case of any original measure, the test of validity should be whether these two characteristics are present to an adequate degree. It is, therefore, a matter of judgment.

The problems involved in developing criterion measures which will be pertinent and comprehensive and which will be acceptable to all interested parties are those of (1) careful definition of the area with respect to which the performance of individuals is to be evaluated, and (2) reduction of the influence of extraneous variables. These problems will be considered in a later section.

Discrimination and Reliability. A criterion measure cannot be considered acceptable unless it indicates, with satisfactory precision, existing differences between individuals with respect to the performance under scrutiny. In general, discriminable differences in success do exist; the measuring technique must not be so gross as to obscure them unduly. Rating scales which permit of nearly all individuals rated being placed in the upper 10 or 20 per cent of the total population illustrate a type of instrument lacking in discrimination.

Not only must a criterion discriminate between individuals who vary in quality of performance, but it must do so reliably. Reliability, in this context, refers to consistency of measurement. If a measure is reliable, the amount of chance error in the assignment of an individual to a position on the scale of success will be minimized. These chance errors can be considered as arising from variations in the individual himself, in the measuring procedures, and in the scoring procedures. Repeated application of a reliable measuring procedure by the same, or by different observers, to a given sample of individuals will yield substantial agreement as to the rank order of merit assigned.

Relation of Discrimination and Reliability to Criterion Acceptability. Discrimination and reliability are both related to the problem of accuracy of measurement (discussed later in this chapter). In relation to a criterion, both pertinence and comprehensiveness are logical concepts, but discrimination and reliability are usually expressed as statistical concepts with the adequacy of each

defined in terms of the criterion measurement data. Discrimination and reliability are interrelated to a considerable extent, but are sufficiently distinct to warrant separate consideration. Discrimination is concerned with the *number of categories of success* to which individuals are assigned, the categories representing observed differences in performance. Reliability is concerned with the *stability* of the values assigned by the measuring technique. It is clear that no matter how much apparent sensitivity in indicating individual differences an instrument may possess, its effective discrimination between individuals is limited by the amount of chance error contained in the values assigned.

The importance of a reliable measure can be seen from the following example. If commanding officer "A" states, in filling out Ensign Smith's fitness report, that he would "particularly desire to have him" and the executive officer, having observed Ensign Smith over the same period of time and under the same circumstances, should state that he "prefers not to have him," the two reports would be in disagreement and the results would not be termed reliable.

If both the commanding officer and the executive officer indicated that they would "particularly desire to have" all of their officers, that one was no better than another in general value to the ship, the ratings would indicate no *discrimination* between the officers being rated and would be worthless as criterion values even though agreement was secured. Of course, rare instances may occur in which a group of officers are actually equally good or poor in all areas of performance under consideration, but the frequent lumping together of individuals on rating forms is almost entirely a failure to indicate differences in proficiency which actually exist.

COMPARABILITY AND CRITERION MEANING. The importance of comparability of values is readily appreciated when comparisons are to be made between individuals who have been tested or rated at different times or under different circumstances, or when data for samples from separate ships, schools, and so forth are to be combined into a total sample. If Rogers, Gunner's Mate third class, is assigned a criterion value of 8 representing a specified level of competence on one ship, and if Jones, Seaman first class, is equally competent in the area being measured, he too should be assigned the same value of 8 even if he is a non-rated man and on another ship. Most scores from test situations and the values from some types of rating devices or check lists can be considered comparable in this sense. Lack of comparability of measures usually develops from the absence of a single set of reference meanings which can be applied to the evaluation of performances.

Since it may be desirable to combine the criterion values for

several groups of individuals for purposes of analysis, and since one usually wishes to generalize from a sample to a total population, the comparability of the values from group to group must often be considered. The use of procedures which yield values comparable only for the individuals in a particular sub-sample of the total population studied, leads to the difficult problem of securing comparability. Statistical techniques of "within-groups" analysis and of adjusting differences in mean scores are available when comparability of values from group to group cannot be assumed; but use of these techniques, especially if the assumptions are poorly satisfied, may result in some loss of precision. It is preferable to employ variables or sets of variables whose values will have a standard meaning irrespective of the individual to whom applied, the group of which he happens to be a member, or the test or rater used to evaluate his performance.

Problems Met in Developing an Acceptable Criterion

The general problems encountered in the development of criterion measures have already been indicated as follows:

Definition of the areas of performance,
Reducing the effect of extraneous variables,
Increasing the accuracy of measurement,
Increasing comparability of test scores.

One additional problem is that of combining measures into an overall composite when several aspects or dimensions of success have been measured separately. This section will present a more detailed discussion of these several problems, together with suggestions for their solution. It should be emphasized again that the problems are interrelated, even though they are divided into certain categories for convenience of exposition.

DEFINING THE AREAS OF PERFORMANCE. Success in Navy billets has many different aspects—technical competence, leadership ability, dependability, overall value to the ship, and so forth. Each of these aspects of success might be regarded as a possible area to be measured. In a particular research study, however, it would be necessary not only to select one or more of them as the area of performance to be considered but also to give a precise definition to those selected. If two or more areas were selected, it would also be necessary to know if they really constituted different facets of success.

DEFINITION IN TERMS OF THE PURPOSE OF THE PROJECT. The particular aspect or aspects of performance to be included in the criterion depend primarily on the purpose of the project. For nearly all

Navy follow-up studies, this is determination of vocational success in performance on the job.

A definition of the purpose of the study may be the only definition of success required, if the purpose can be precisely stated. For example, if the purpose of a study is to predict the success of men copying both machine and hand-sent radio code under operating conditions, no other definition of the criterion is required. This statement delimits the areas of success to those pertinent to the purpose of the project. A similar restriction in the areas to be included would occur if the purpose of the project was to evaluate the effectiveness of a motor machinist's mate school in teaching the operation, maintenance, and repair of diesel engines. In this situation, an individual's leadership qualities, his adjustment to Navy life, and his work habits would be irrelevant except insofar as they might be germane to success in operating, maintaining, and repairing diesel engines. If developing technical competence were the prime objective of the school, the criterion used to evaluate the school's effectiveness should be confined to measures of technical competence on the job.

On the other hand, the objective of a research study may be a very general one, such as the prediction of overall success as a signalman. This objective will require the evaluation of performance in all areas which make for success in this billet. Success as a signalman may include proficiency in receiving and sending flashing light messages, knowledge of communication procedure, skill in sending semaphore or in bending on and hoisting signal flags, ability to work with and handle men, proficiency in standing lookout watches, and the satisfaction expressed for the job.

The purpose of the project sometimes can be presented in the form of a functional question which will delimit, by implication, the areas of success. Such a functional question was used in a study of the performance of junior officers aboard ship. Each reporting senior officer was asked to nominate the junior officer he would most want to take with him to his next ship (the purpose of the study being to identify superior and inferior junior officers). This statement was sufficiently meaningful to these senior officers. They tended to agree on the officers so selected, even, in some cases, to the extent of giving identical reasons for wanting or not wanting a specific junior officer. The nominations for high or low, since they represented reasonably reliable ratings, constituted data by means of which a set of criterion groups was identified.

Selection of Important Areas of Performance. If analysis of the purpose of the project indicates that several areas of performance would be included in the criterion, these areas can be identified

through (1) job analyses which include statements of the psychological, social, and physical characteristics of the job, (2) opinions of experts as to the importance and relative frequency of occurrence of the suggested areas, and (3) experimental and statistical studies of proposed criterion measures. Complicating factors in identifying the important areas are the problems of securing job analyses which are more than billet descriptions, of selecting experts who are competent to make the necessary judgments, of insuring that minimum essentials are included even though some groups are homogeneous with respect to such areas, and of submitting the measures to an experimental check.

Differentiating Between Areas of Performance. If a measure is to be used as indicating the success variance in only one aspect of performance, it should be possible to show that the measure does reflect success in that area alone. This requirement means that the values on one criterion measure must be to a discernible extent independent of the values representing success in other areas. Judgments of "technical competence" and of "petty officer qualities," for example, were found in a study reported in Chapter XX to be practically the same judgments even though extensive definitions of each of these two areas had been set up, and even though in isolated cases, instances of a lack of correlation between the two aspects of success aboard ship were found. In the light of these data the terms "technical competence" and "petty officer qualities" do not appear to indicate aspects of success which can be differentiated by rating procedures.

The high average intercorrelation usually secured between ratings of a wide variety of traits may mean that the traits as found within the individuals are actually highly correlated, or it may mean only that the judges are unable to discriminate between these traits because of a general "halo effect." To ask raters to make separate judgments of a group of individuals on four or more traits will usually prove to be asking them to give the same evaluation four or more times. If such traits are to be differentiated, other techniques, such as tests, must be developed for their differentiation and measurement.

The evidence that the areas subsumed in a criterion are differentiable is statistical in nature if values for the areas in question can be assigned. A comparison of the correlation coefficient between two variables with the geometric mean of the reliabilities of the same two variables will indicate roughly whether the two measures can be considered as measures of the same or of different areas. A better test of the hypothesis that several dimensions are involved may be secured from a factorial analysis of a set of criterion measures, the

results of which will indicate the number of dimensions required, the variables which can be used to measure the different dimensions, and the extent to which the common factors account for the variance of the separate measures. If a single overall measure has been secured, the question of what is being measured can only be answered with descriptions of the procedures used and of the data obtained. The research worker must then decide on the basis of these descriptions whether the results can be accepted as indicating performance differences in the defined area of success.

ADEQUACY OF DEFINITION OF AREAS OF PERFORMANCE. Satisfactory evaluation of the adequacy of the definition of success can only be arrived at after preliminary trials of the proposed measures. Necessary for such evaluation are (1) descriptions of procedures used (list of experts consulted, outline of job analysis results, statement of operations used to increase the probability that the important areas are covered and that the effect of extraneous factors has been considered), and (2) statistical data showing the extent to which success in several areas can be differentiated, the discrimination obtained between individuals, and the reliability of the measures. Since these data also constitute evidence of accuracy of measurement, the adequacy of the definition of the areas cannot be satisfactorily evaluated until the problems of measurement (discussed later) are also solved.

Reducing the Effect of Extraneous Factors

Even though the areas of performance comprising the criterion measure have been carefully defined, there is always the danger that extraneous factors will contaminate the final results by introducing irrelevant sources of variation. Some of these factors exist in the environment in which the individual works, some may be inherent in the equipment which he operates, and others may exist in the minds of those who do the evaluating. In any case, every precaution should be taken to reduce the effect of these factors to a minimum.

CONTAMINATION BY KNOWLEDGE OF PREDICTOR VARIABLES. One such extraneous factor is the knowledge and use of selection and training data in the rating for advancement of individuals on the job. Knowledge of the background information results in a contamination of the criterion, since an individual's position on the scale of success is no longer a function solely of success on the job. The use of this background information is generally defended on the ground that increased precision in promoting individuals or in assigning them to specific jobs is achieved through such knowledge. It should be pointed out, however, that this does lead to a spuriously

high relationship between the selection or training variables and the criterion. For the most acceptable criterion, this type of contamination of the criterion data should, if not excluded, at least be considered in evaluating the results. (See Chapters IX and XI for illustrations of the operation of this factor.)

BIASES OF JUDGES. Both supervisors and peers frequently possess misconceptions of the importance of particular characteristics of behavior. Their judgments reflect not variations in performance but the judges' likes or prejudices. The military rank held by the individual, his length of service in the organization, and his total amount of experience in the Navy are illustrative of the variables which may cause a rater to err in his evaluation of an individual's level of competence. The preconceptions held by the respondents regarding the desirability of transferring individuals from one type of Navy duty to another or of the desirability of shipboard as compared with school training are also sources of bias. These preconceptions can sometimes be evaluated, or their effects can be lessened, provided the procedures used in training of raters and in the collection of data are sufficiently extensive and complete.

EFFECT OF AMOUNT OF EXPERIENCE. The effect of amount of experience on the job upon criterion values should be estimated, because the criterion value desired should represent the position of the individual at some fixed point of experience. There may be some value in developing a measure which will reflect amount of experience if such a measure is not directly available, but amount of experience can usually be determined directly from personnel records. The effect of experience will be apparent in objective measures of performance as well as in ratings. This effect may be partialled out statistically from the criterion measure, or the study can be limited to individuals who do not differ significantly in their amount of experience. The result of the application of the statistical procedures will be a set of criterion measures which provide estimates of the relative positions the individuals would have had if all possessed the same amount of experience. (See Chapter XX for a practical solution of this problem.)

CHANCE ASSIGNMENT FACTORS. Another type of extraneous factor is that which results in the assignment of an individual to an older ship instead of to a newer one or to one billet rather than another. Two individuals, equally competent, may have had marked differences in opportunities for advancement and for increase in skill and general competence. The net result of this inequality is that they will be evaluated by superiors or peers on quite different standards; their performance on objective measures may also be quite different. The criterion measure will reflect these differences in

performance but the source of the differences cannot be traced to the usual predictive factors. It appears to be true that replacements sent to a well organized ship with a fairly full complement will have fewer opportunities for learning all aspects of given jobs and for showing their abilities and skills than if they had been sent to a newer ship. Similar chance factors may result in the individual's being assigned to a job with the ship which offers much or little opportunity for advancement. The evaluation of the effect of some of these factors upon the criterion can sometimes be accomplished through the design of the sampling procedures, thereby increasing the precision of the results.

Increasing the Accuracy of Measurement

Associated with the problems of restricting the areas of performance to those pertinent to success are the problems of assigning values which can be used to represent differences in performance in these areas. The process of assigning the appropriate values, usually called measuring, consists of securing judgments (of "greater than or less than," "present or absent," and so forth) with respect to some aspect of an individual's behavior, including his performance in controlled or test situations. Attempting to increase the accuracy with which these values are assigned raises such questions as: (1) Are the procedures used such that similar values will be assigned to an individual's performance by comparable judges (or tests) and at different times? (2) Will the values assigned permit the separation of the individuals into groups homogeneous with respect to judged differences in behavior? The first question deals with problems of reliability of the procedures; the second question is concerned with securing measures which discriminate between individuals. It should be emphasized that these two questions do not deal with the problem of what is being measured, i.e., whether the values on different measures represent differences in behavior in separately defined areas of performance. Only if the several measures prove to be reliable and permit discriminations between individuals, can the question of whether the values represent differences in differentiable areas be raised and determined.

Problem of Reliability. Improvement in reliability is concerned with reducing the amount of random fluctuation or chance error that may occur in the measurement process. Close agreement (between comparable tests or independent judgments of respondents) as represented by correlation coefficients of .80 to .90 can be considered as acceptable criterion reliability. Lack of agreement, or low reliability, is often traced to three sources: (1) characteristics of the

measuring procedures, expressed as "test reliability," (2) the nature of the evaluation procedures, termed objectivity and expressed as "reader reliability," and (3) the stability of the trait within the individual himself, which is expressed as "individual reliability." These three sources of error variance are discussed in any adequate text on psychological measurement, but they are mentioned here because of their importance in developing an acceptable criterion. The operation of any of these sources of error will lead to the same statistical result—a lack of agreement.

Reliability of Test Situations. In carefully controlled test situations the resulting scores are usually more reliable than those secured from rating procedures. The lack of reader reliability in the assigning of scores is usually effectively reduced by making the test objective, i.e., by securing responses in such a way that the personal biases or preferences of the scorer are minimized or eliminated from the evaluation of the performance. The use of devices such as quality scales, short answer written items (as multiple choice or matching items), accurate tolerance gages, and check lists for performance tests may result in close agreement between the scores or values assigned by two or more readers, checkers, or scorers.

The lack of test reliability can often be traced to one or more of the following characteristics:

(1) the inclusion in the test of an inadequate sample of the individual's proficiency,
(2) ambiguity in the directions or test situations,
(3) the individual's unfamiliarity with the test situation,
(4) differential motivation or acceptance of the test situation by individuals,
(5) variation in the testing conditions or procedures,
(6) variation in the difficulty level of the parts of the test, and
(7) variation in the homogeneity of the tasks covered by the test.

The first two items are those most frequently in need of improvement. A more adequate sample of performance may require longer, more carefully designed tests, or it may require a set of tests covering the various sub-areas. Ambiguous directions or tasks will not mean the same thing to different individuals, and a given person may well interpret them differently on two separate occasions. The solution of these problems is often a trial and retrial process.

Reliability of Rating Situations. Estimates of the reliability of ratings and rankings tend to be lower than those obtained in test situations. This general increase in the estimated error variance of judgmental values is traceable to an appreciable degree to the difficulty of securing comparable judges. In a test situation comparable

forms of the test are defined in terms of the selected content and in terms of the statistics of the test-score and item-response distributions. Comparable judges, however, are sometimes defined (in a circular fashion after the ratings have been obtained) as those for whom agreement is secured.

Comparability of judges should be discussed first in terms of the opportunities each judge has had to observe the characteristics to be rated, his ability to make the desired discriminations, and his interest in doing so. But securing comparable judges does not in itself insure reliable ratings. Variation in the relative weights by which the judges combine their impressions, variations in the behavior observed and in the way it was observed, and variations in the standards of evaluation from one judge to another lead in general to disagreement between judges.

Extensive training of the judges in eliminating personal biases, in evaluating a certain pattern of behavior, and in weighting the elements in a similar fashion is one means of increasing the reliability of the ratings. The use, both during the training period and during the collection of data, of extensive check lists and descriptions of the trait rated (expressed in terms of the respondent's vocabulary) tends to reduce variations in interpretations. Selecting judges who have had an opportunity to observe the behavior over a period of time, training them, requesting only judgments upon discriminable characteristics, and directing the process of rating or ranking in a personal interview as was done in the collection of criterion data discussed in Chapter XX should lead to judgmental values of acceptable reliability.

Discriminating Measures. While reliability is an important attribute of a measuring instrument, it is equally important that the measure indicate differences between individuals, differences which reflect the degree of success achieved. Solutions to the problems of measuring a defined area of success, and of measuring it reliably, will be of little value unless the criterion measure provides a spread of scores so that the more capable individuals can be effectively separated from the less capable. Homogeneous groups can be found in which the individual differences on the proposed measure of performance are so small that they cannot be reliably discriminated, with the result that all of the individuals in such groups are assigned the same criterion value. In such a situation more discriminating measures must be secured if the criterion is to be considered acceptable.

Discriminable differences in success do exist even in homogeneous groups and in areas of performance which can be called "minimum" essentials. The problem is to develop, within the defined areas, tech-

niques which will insure that observable differences will be indicated and that the assigned scores or ratings do provide a spread of scores. A spread of judgments or ratings can be secured by using order-of-merit rankings or by requiring the judges to place a given number of individuals in each category of the rating scale. The ratings or rankings from a group of judges can then be scaled by available psychometric procedures.[1] If the ratings have been secured by a procedure recommended by Horst,[2] and providing the judges have rated the same men, an origin and unit can be determined for each judge. Discriminating measures can usually be secured from test situations by varying the difficulty and number of tasks presented or by using length of performance time as well as accuracy in the scoring. Unfortunately, test scores or grades, as well as ratings, often exhibit a lack of discrimination so marked that extensive research is necessary to develop a suitable technique for discriminating between levels of performance.

Increasing the Comparability of Criterion Values

Since the absence of a rational zero point and a rational unit is characteristic of nearly all psychological measures, the problem of comparability of the values or categories used in measures of performance is a general one. Comparability refers to the constancy of interpretation, i.e., will the assigned values represent approximately equivalent levels or characteristics of performance for different individuals, different groups, and different situations? The procedures suggested to increase the comparability of the criterion values will differ somewhat, depending upon whether tests or the judgments of observers are used to obtain the criterion measures. A related problem is concerned with the units in which the levels of achievement are expressed.

TEST SCORES. If the criterion values are determined in a single carefully controlled test situation, as for example, from a standardized written achievement test of the multiple-choice type, the problem of comparability of values from this one measure is, for practical purposes, solved. An individual who secures a score of 12 on this type of controlled measuring situation on one ship can be considered to have achieved as many items as another individual on a different ship who also is assigned a score of 12 on the same testing situation. Such a score does not tell how much better one individual is than another or what a score of this magnitude means in terms of

[1] Guilford, J. P. *Psychometric Methods,* Part II. McGraw-Hill, New York, 1936.

[2] Horst, Paul. *The Prediction of Personal Adjustment,* pp. 76-78. Bulletin 48, Social Science Research Council, New York, 1941.

level of performance. The latter could be accomplished only by some sort of scaling procedure.

RATING VALUES. A much more difficult problem of comparability occurs when no test techniques are applicable or feasible and when a criterion consisting of ratings is used. Noncomparability of ratings may result from variations in (1) the judges' interpretations of the definition of the scale, (2) opportunities for observation of the behavior being judged, (3) the individual standards applied to the observations made, and (4) the abilities of the judges to make the desired discriminations. The procedures (described earlier) of training the raters and of confining the ratings to observable behavior are not only means of improving the reliability of judgments but also of improving the comparability of the ratings. The requirement of a high level of agreement between judges for the assignment of individuals to categories on a qualitative variable provides a means of increasing the assurance that the judgments do represent comparable performance characteristics.

If the ratings purport to represent the positions of individuals on a scale of success, as do order-of-merit rankings and various types of rating scales, lack of comparability most frequently arises from differences in the standards of evaluation used by different judges. Some judges consider a particular characteristic as being more, or less, damaging to overall success than do other judges. These differences in standards are also complicated by differences in the level of achievement which are characteristic of the groups the judges have known in the past. Attempts to have the judges consider "all the men in the Navy today" have not given results much different from those secured from comparisons confined to the groups actually being judged. Differences in the means and standard deviations of the ratings of trained judges, however, can be made small enough to permit the acceptance of the rating values as comparable.

The use of an order-of-merit ranking or the use of ratings in which the minimum number of individuals in each category is specified provides a means of forcing discriminations between individuals which, other things being equal, will increase the comparability of the ratings. If the ratings or rankings by a number of judges on a trait, or on a set of traits, are secured for a group of individuals, the differences in the means and standard deviations of the ratings given by individual judges can be adjusted, and more nearly comparable scores can be obtained by Horst's [3] procedure. These procedures, however, do not provide a means of comparing individuals from group to group unless there exists an appreciable overlapping of respondents between groups. Since this condition rarely obtains

[3] Horst, *loc. cit.*

in the Navy, comparability of ratings from one group to another is often assumed on the premise that the means and variances of the several groups would differ only by chance. When this assumption is made, order-of-merit rankings are often used and the ranks converted to standard score or sigma values on the further assumption that the distribution of the true measures of performance is normal. When the values can be considered as representing only an individual's position in an intact group and the mean criterion values are not assumed to be equal, only within-group analyses of the criterion scores are permissible.

Another procedure for increasing the comparability of ratings requires the use of several criterion measures, including at least one measure correlated with the ratings which can be assumed to result in comparable scores from group to group. Differences in the rating means and standard deviations of the groups can then be assumed to be proportional to the differences in test score means and standard deviations, and appropriate translation equations can be determined.

Units of Measurement. When criterion measures are to be combined, both the constancy of meaning of scores on a single measure and the extent to which similar criterion symbols represent similar levels of achievement should be considered. If the set of measures consists of scores from controlled test situations, the constancy of any one measure can generally be assumed. The problem then involves developing a set of standards for evaluating the performance on all the different measures. These standards are usually secured by establishing sets of derived scores, such as standard scores or percentile scores, in which the separate performances are evaluated in terms of the performances of some standard or norm group. The use of such derived scores provides an arbitrary reference point, for example a mean of 50, and an assumed unit for all of the measures, such as is provided by a scale with a standard deviation of 10. The combination of such derived scores then is not influenced by the variations in the standard deviations and in the means of the measures, which are usually artifacts of the test design.

When judgmental ratings are to be combined with test scores, or when different ratings are to be combined, adjusting the means and variances by using standard scores is also appropriate. The original ratings, however, should provide a fairly adequate spread of scores and should be reliable; otherwise, the combination may give disproportionate weight to the unreliable and non-discriminating measure. If several judges have been used, the differences in means and variances of their ratings should be adjusted before combining the values.

Combining Criterion Measures

It has been pointed out that in some situations the requirement of comprehensiveness cannot be met by any single criterion measure. In other circumstances, judgments of overall success cannot be secured reliably or with sufficient discrimination. In still others an overall criterion may be unobtainable, or obtainable only at prohibitive cost of time. In many of these cases the difficulty of securing an acceptable criterion can be solved by developing sets of measures, each unit of which satisfies the requirements of pertinence, reliability, and discrimination, and then combining these separate measures to form a single composite criterion. In the procedures for combining measures which are discussed below, two general cases occur: (1) an overall measure is temporarily available for use in the research, but it must be replaced by a group of measures in a practical situation; (2) no overall measure is available, even for the research study.

COMBINING MEASURES USING AN OVERALL CRITERION. Whenever an overall criterion is available, it may be used as the controlling variable for evaluating and weighting a set of supplementary criterion measures to form a composite variable. The problem is to combine and weight the independent variables so that the composite will conform as closely as possible to the overall values with the desired increase in reliability, discrimination, or convenience.

If the overall criterion is a quantitative one, the procedure to be used is basically a multiple regression procedure in which the dependent variable is the overall criterion. Consider an example of this type of problem and its solution by one of the services during the war. It was desired to develop a criterion of officer success which could be routinely applied in peacetime as well as war. The unique conditions of war provided a large group of officers who were capable of judging the overall performance of their fellows in combatant and non-combatant situations. By pooling the judgments of these officers as to the success achieved by their fellow officers, an acceptable criterion of overall performance was secured; but this criterion was not one that could be secured except under these special war conditions. Therefore, in order to secure a more generally applicable criterion, a series of supplementary measures which could be applied at different times and under different conditions was developed simultaneously with the overall judgmental criterion. These supplementary measures consisted of achievement test scores, ratings by supervising officers on carefully constructed rating forms, and ratings by panels of selected officers. These supplementary measures were then combined so that the single composite value

corresponded as closely as possible to the averaged overall judgment of the large group of fellow officers. This procedure of combining criterion measures can be used only when a single measure is available to determine the weights to be assigned other measures in combination.

If the overall criterion is a qualitative one (e.g., provides only criterion groups characterized by the presence or absence of some attribute), the problem is still one of determining the weights which, when applied to the supplementary measures, will differentiate between the two groups as much as possible. This problem can be solved through the use of Fisher's discriminant function, or a suitable approximation of it. It has been shown that the discriminant function can be considered as a regression problem with the overall criterion, the dependent variable, assigned values of 0 and 1.[4]

The most crucial phase in the procedure described in the preceding paragraphs concerns the acceptability of the overall criterion, since it is used to determine the proper weights to be assigned the supplementary variables. If the overall criterion is biased, or if it unintentionally places an emphasis on relatively unimportant variables, or if some variables are omitted entirely, the composite will reflect these same weaknesses or biases. When the overall criterion is used as the controlling element in selecting supplementary measures, the closest scrutiny of the suitability of this measure is required.

Other complicating factors in the combination of variables are the reliabilities of the separate measures, the units in which they are expressed, the homogeneity or heterogeneity of the sample population, and the extent to which the supplementary measures cover all of the important areas of success, that is, the comprehensiveness of the set of measures.

The influence of restriction or of pre-selection of the population on the covariance matrix has been noted in Chapters XII and XIII. The effect on the intercorrelations of the separate criterion measures is similar to the effect on sets of prediction variables, the net result being to depress the magnitude of the average intercorrelation. As long as the measures are to be used with the sample population or with similar populations, the obtained coefficients can be used. If the measures for a more heterogeneous population are to be combined, using the weights determined from the homogeneous sample, estimates of the population weights might be made using the procedure suggested by Burt,[5] in which the sample regression weights

[4] Garrett, H. E. "The Discriminant Function and Its Use in Psychology," *Psychometrika*, vol. 8, No. 2, p. 65, June, 1943.

[5] Burt, C. "Validating Tests for Personnel Selection," *British Journal of Psychology*, vol. 34, Part I, p. 9, September, 1943.

are multiplied by the ratio of the standard deviation of the variable in the population to the standard deviation of the variable in the sample.

COMBINING MEASURES WITHOUT USING AN OVERALL CRITERION. Some situations exist in which the separate criterion measures are to be combined to form a composite variable without using an overall criterion as the dependent variable. This situation may arise either from an unwillingness on the part of the research worker to accept an available overall measure as a suitable criterion or from a lack of any such overall criterion.

An illustration is provided in the formulation of a composite criterion for yeomen. The areas of performance to be measured are proficiency (1) in typing, (2) in taking dictation, and (3) in filing; but none of the supervisors or peers available is sufficiently familiar with the individuals' achievements in all of these areas to make an overall appraisal of the yeomen's proficiency. Each of these areas, however, can be measured reliably by suitable performance tests, and these test scores can then be combined to establish a composite measure of success as a yeoman. The actual procedure would be to multiply each separate score by an appropriate value, or weight, which represents the relative importance of the area, and then to add the products for each individual. The weights to be assigned the separate measures in such cases can be determined by using (1) sets of rational weights based on the judged importance (or frequency of mention) of the separate areas, (2) the technique of maximizing the composite criterion variance (which results in maximizing the internal consistency of the set of criterion measures),[6] and (3) weights based on the standard deviations and the reliabilities of the separate measures.

The first procedure of assigning rational weights in terms of importance involves combining the measures so that the rational weights represent the actual effective contributions of the separate measures to the variance of the composite.[7] The nominal weights, or regression weights, must then be determined so that the effective contribution of the measures will be equal to the rational weights. The solution requires the analysis of a set of simultaneous equations which express the nominal weights in terms of (1) the standard deviations and the intercorrelations of the measures, and (2) the effective or rational weights to be secured.

The proper use of rational weights is an acceptable procedure for combining measures. The chief weaknesses in the procedure are lack of recognition of the differences between effective and nominal

[6] Horst, *op. cit.*, p. 73.

[7] Richardson, M. W. "The Combination of Measures," in Horst, *op. cit.*, p. 383.

weights, and failure to secure competent judges for determining the rational weights. Combining instructors' marks, achievement test scores, and aptitude ratings through the assignment of rational weights is the procedure usually employed for determining final school grades. The usual rational weights for school marks are either unity, so that the marks are averaged, or simple ratios as 1:2 or 1:2:5. Unfortunately, the rational weights often are assigned on the basis of little or no evidence and rarely represent the actual combination.

The weighting system used in one of the reserve midshipmen's schools provides an example of this failure to consider the problem of effective contributions of variables. The instructional staff had decided, by some undefined procedure, that the academic multiple (a weighted average of weekly marks and final examinations) was to be weighted 4 to 1 for the purpose of combining it with a rating of aptitude for service. Since no intercorrelations or total variance contributions had been computed, it can be assumed that the ratio 4:1 represents the judged relative importance of these two sets of values as measures of success in this midshipmen's school. But when the relative contributions of the two measures to the composite were determined, the effective weighting was 17:1. Because the rating on aptitude for service received such a low effective weight, the correlation coefficient between the rating and the composite was only .05, and this low correlation was then considered as evidence for omitting the aptitude for service rating from the criterion.

The second procedure, of maximizing the internal consistency or the dispersion of the composite scores, is computationally difficult if more than three measures are to be combined. It is recommended for use in the case in which the measures can safely be assumed to be measuring the same thing, as would be true of three different radio code receiving achievement tests. The method must be used cautiously since the solution depends upon the units in which the measures are expressed.

If it can be assumed that all measures are measuring essentially the same thing and that the differences in standard deviations are artifacts of the measures, weighting in terms of the reliabilities and standard deviations of the separate measures may be considered. But the generally applicable procedure recommended for the combining of measures without the use of an overall criterion is (1) to develop reliable and discriminating measures of relatively independent areas, (2) to determine the desired rational weights from the pooled judgments of the relative importance of each area, the judgments being obtained from carefully selected competent judges, and (3) to combine the measures so that the effective contribution to the total variance will be in the same ratio as those of the rational weights.

Acceptability of Navy Criterion Measures

The criteria of success employed in Navy schools and in the evaluation of job performance consist of grades or marks, scores on achievement or proficiency tests, and ratings. In the preceding chapters frequent mention has been made of each of these three types of criteria, and some of their defects have been pointed out. In this section each will be evaluated by the standards of an acceptable criterion which have been described in the early part of the chapter.

Grades or Marks. Grades assigned in naval schools are usually based on scores in essay or objective tests, observations of laboratory performance, or subjective evaluations of work produced. In some schools the job of evaluation has been taken seriously; in others it has been regarded as only a necessary evil of school administration.

Grades at their worst have been found to lack all of the qualities of an acceptable criterion. Even in the best schools, grades usually lack one or more of these qualities. It was not uncommon, for example, to find schools giving grades of acceptable reliability which were quite lacking in comparability and comprehensiveness. Since the schools did not receive instructions about how to grade until relatively late in the war, it is not surprising that such an uneven job of evaluation was being done. Nevertheless, it must be admitted that grades in Navy schools as a whole fall considerably short of acceptable standards in measuring the achievement of trainees.

Achievement or Proficiency Tests. During the last year and a half of the war a large number of standardized achievement examinations were introduced in the elementary enlisted schools; a few were developed on the officer level. Advancement examinations were prepared to assist examining boards in the construction of tests which could be used to measure the outcomes of in-service training. These examinations have been described in Chapters XV, XVI, XVII, and XVIII.

Such achievement examinations were usually quite reliable, discriminated well between levels of performance, yielded comparable results from school to school, and measured knowledge or skills pertinent to the school's objectives. Perhaps the chief question which could be raised about these examinations is whether they were comprehensive enough. While an effort was made to build performance and identification tests wherever needed, the majority of tests were of the paper-pencil type and almost certainly did not measure some of the skills which the schools aimed to teach. In addition it should be said that these examinations measured technical competence only and not the petty officer qualities often regarded as very important in determining the success of the man on the job.

The increase in the acceptability of school grades as criteria of success through the use of standardized achievement examinations has been pointed out. In a few cases studied, the use of such examinations has resulted in an increase, or a shift, in the predictive efficiency of the tests of the Basic Test Battery. These changes in validity coefficients probably can be traced to the improved definition of the objective being measured, to the increased reliability of the criterion, and to reduction in the effect of extraneous factors.

The effect of using *Advancement Examinations* as measures of job proficiency has not been studied. Presumably these examinations will improve the quality of the technical examinations constructed by the examining boards; and when used in conjunction with more objective Practical Factor Tests, they should provide improved measures of the job proficiency of enlisted personnel. Since the trade testing approach has been used by the Navy for many years in advancing enlisted personnel from one pay grade to another, the problem of improving the measurement of technical competence is primarily one of improvement of testing techniques and of clarification of the areas of performance to be included in the definition of success.

A noteworthy study of the performance of destroyer radiomen, conducted by one of the fleet commands, utilized a series of performance and written examinations administered by teams of communication specialists. The results of this project indicated the value of having accurate measures of performance and suggest that the continuation of such studies would be desirable. The acceptability of properly prepared achievement examinations as criterion measures will depend largely upon the pertinence and comprehensiveness of the measures, since the technical problems of reliability, discrimination, and comparability are usually satisfactorily solved in such measures.

RATINGS. The evaluation of the job performance of subordinates is routinely made in the Navy by tens of thousands of judges, both officers and enlisted men. Enlisted men are given marks, either quarterly or on change of duty, for technical proficiency, seamanship, ability to lead men, and conduct. Fitness reports, filled out periodically for all officers, include ratings on the officer's qualifications and overall performance in his present duties, on his initiative and responsibility, on his understanding and skill, on leadership, and on his conduct and work habits.

Unfortunately there is considerable doubt as to whether these marks are assigned on a uniform basis and whether the ratings on these separate categories represent anything more than slight modifications of the overall judgments. A great proportion of officers and

petty officers have never received specific instruction in how to rate performance; consequently it is not surprising that the ratings in general fail to meet the characteristics of acceptable criterion measures.

Statistical studies of the ratings indicate that they are lacking in discrimination, reliability, and comparability. Especially noticeable are the facts that most ratings are high and that the means and standard deviations of the ratings assigned by the many different judges vary considerably. Although there is less statistical evidence on the pertinence and comprehensiveness of the ratings, there are indications that they are based heavily on observations of personal qualities and less on technical competence. Because of numerous deficiencies, the present ratings of quality of job performance do not constitute acceptable criteria of success for use in research studies.

Summary

The characteristics of an acceptable criterion, the problems encountered in the development of such a criterion, and an evaluation of criteria used in the Navy have been presented in this chapter. This discussion has indicated that there is no simple solution to the problem of developing an acceptable criterion; procedures which in one situation lead to an acceptable criterion may not result in acceptable measures in another. The development of an acceptable criterion is, in actual practice, accomplished by a series of successive approximations. First, there is usually an attempt made to secure a reliable measure of individual differences in a broadly defined area of performance. Next, the definition of what constitutes success is clarified, stated more precisely, and new studies are made. As a result, improvements in the characteristics of the criterion may be effected over a period of time. This process may be continued until an instrument is developed which reflects with a satisfactory level of acceptability how well the individuals measured actually perform on the job.

The problems described in this chapter are nowhere more acute than in the development of an acceptable criterion of shipboard performance. In such a criterion the variations in the type of population, in the nature of work done, and in the standards of success encountered tend to be maximized. The problems of definition, of accuracy of measurement, and of comparability of units are particularly troublesome. The problems actually encountered in the development of a criterion of shipboard performance, and the techniques used in attempting to solve some of these problems constitute a major topic of the next chapter.

CHAPTER XX

PREDICTION OF PERFORMANCE OF ENLISTED PERSONNEL ABOARD SHIP

It is generally agreed among personnel research workers that the validity of prediction variables should be investigated by correlating them with success on the job. In the Navy during the war it was necessary to limit validity studies almost exclusively to determining the relationship between prediction variables and success in training. The results presented in earlier chapters indicate that valuable information concerning the usefulness of classification data were obtained in this manner. To the extent that success in training is related to job success, these studies showed indirectly the relationship between prediction variables and performance on the job. There was a strong desire, however, on the part of those concerned with the improvement of classification and training procedures to conduct studies showing the relationship between various classification and training data and performance aboard ship. In this chapter are presented the results of a pilot study conducted to determine the relationship between certain data on the Enlisted Personnel Qualifications Card and the quality of performance aboard ship of men in six enlisted ratings.

Nature of Shipboard Performance. The complexity of the problem of predicting success in performance aboard ship can be better explained by presenting a brief description of the nature of shipboard duties. In some respects predicting success in the job aboard ship is very similar to predicting successful job performance in civilian life; in other respects the Navy situation is very different from that in civilian occupational life. A machinist's mate, for example, operates, maintains and repairs steam engines much as a civilian engineer does in a power plant. A senior petty officer instructs apprentices (Navy strikers), lays out repair jobs, and sees that necessary checks on operating equipment are made, in much the same manner that a foreman does. Two chief differences between Navy and civilian job performance are (1) the living and working together in close quarters aboard ship and (2) the added stress placed upon Navy personnel by participation in combat. Because of these two factors, supervising officers in the Navy place considerable emphasis upon such personal qualities as ability to get along with others, faithfulness, dependability, willingness to take orders, and interest in the job. A crack radio operator, for example, who is thoroughly disliked by his shipmates may exert a detrimental influence on the whole team; on the

other hand an average operator who is well liked may be a source of inspiration and encouragement to the others to do their best; thus he increases the efficiency of the team.

Population Studied

The sources from which a particular ship may receive men are similar to those for advanced service schools. The men may be sent to a ship direct from a recruit training station, from elementary or advanced schools, or from another ship. As a result of these multiple sources, the men in a single rating group can be expected to be quite heterogeneous with respect to amount of Navy experience, extent of civilian education and of civilian occupational experience, age, and scores on the Basic Test Battery. Since experience can be expected to influence an individual's performance, it would be desirable to have each sample limited to men of a given amount of Navy job experience (or in the rating, as it is termed). Because the rating groups available for this study were so small, it was necessary to include nearly all available cases. The only limitation imposed was that neither men of less than six months shipboard experience on the job nor those of more than fifty months experience in rating were to be included. Since this range of experience included the larger percentage of men in the Navy during the war, it is considered that this sample is sufficiently representative of the Navy population with respect to length of experience.

Sample Design. In order to include in this pilot study the major sources of variation which might influence performance in shipboard duties, the sampling design was made to include ships of different types, men in different pay grades within a given rating, and several different ratings. Six parallel studies, each concerned with a specific rating, were therefore conducted simultaneously in three types of ships. The six ratings, radioman (RM), signalman (SM), radar operator (RdM), fire controlman (FC), machinist's mate (MM), and gunner's mate (GM), were selected so as to sample both deck and engine room, both operating and maintenance ratings. The three types of ships chosen were destroyers, carriers, and large ships (two types of cruisers, and one battleship). For convenience the latter will simply be referred to as cruisers (designated CL) since only one rating group was studied aboard a battleship.

Rating Groups Used. A total of 120 groups from 27 ships (9 destroyers, 12 carriers, and 6 cruisers) was studied. The groups in each rating were distributed as follows: 21 groups of radiomen, 15 groups of signalmen, 22 groups of radar operators, 14 groups of fire controlmen, 29 groups of machinist's mates, and 19 groups of

gunner's mates. Each rating had at least two groups in each ship type, except that no fire controlman or gunner's mate groups were secured from the carriers. Because intact experienced groups had to be studied, the number of men in each group was small, the median number of cases per group ranging from 15 for signalmen to 18 for gunner's mates. The actual numbers in the groups ranged from 6 signalmen on a destroyer to 33 radar operators on a cruiser. The number of ships by type, number of cases by rating, and total number of cases included in the study are shown in Table 1-xx. The ships used were those which (1) were available in West Coast ports between June and August 1945, (2) had engaged in combat or

TABLE 1-XX. Distribution of sample population by ship type and by rating

Type of Ship	Number of Ships	Number of Cases	Number of Cases by Rating					
			RM	SM	RdM	FC	MM	GM
Destroyer	9	460	72	12	77	75	87	137
Carrier	12	682	172	104	194		212	
Cruiser	6	726	122	82	94	144	128	156
All ships	27	1,868	366	198	365	219	427	293

forward area operations, and (3) had available personnel records of some appreciable degree of completeness. The sample was not strictly representative of combatant vessels of the fleet at large in terms of the relative number of ships of each type and of different dates of commissioning. Fewer destroyers and more newer ships were included (commissioned after December 1943) than would be in a representative sample. However, it is considered that the data should be indicative of the results which would be found on a large number of combatant ships in the Navy in the six ratings covered in this study.

Predictive Factors Studied

Performance aboard ship can be expected to be influenced by a number of variables. These factors may conveniently be grouped into two categories (1) those not associated with Navy experience, and (2) those associated with Navy activities. The first group includes such variables as an individual's personal qualities and abilities, his civilian experience, and his training. The second group of variables has to do with factors which represent gains in proficiency, poise, and prestige as a result of a tour of duty in the Navy. The factors studied in each of the two categories are presented below.

FACTORS NOT ASSOCIATED WITH NAVY EXPERIENCE. These variables may be summarized as follows:

1. *Scores on the following tests of the Basic Test Battery:* General Classification Test (GCT), Reading (READ), Arithmetic (ARI), Mechanical Aptitude Test (MAT), Mechanical Knowledge Test (Mechanical Score) (MKM), and Mechanical Knowledge Test (Electrical Score) (MKE). The special tests of the Basic Test Battery were not included because an insufficient number of scores was available.
2. *Age,* as given by year of birth. This factor was included because for a number of ratings, age qualifications are specified.
3. *Civilian education,* as defined by highest grade completed. The amount of civilian schooling is a variable which has frequently been listed as one of the qualifications for success in particular ratings.
4. *Civilian occupation,* classified as managerial, clerical, mechanical, or miscellaneous with two levels of amount of experience in each category—level 1, one to four years of experience; level 2, more than four years of experience. Less than one year of experience in any job was listed as a separate level.
5. *Classification interviewers' judgments,* as reflected in the assignment of quality classification codes and elementary school recommendations (if made at the recruit training station). These have been described in Chapter III.

FACTORS ASSOCIATED WITH NAVY EXPERIENCE. Besides the variables listed above, the study included a number of other factors which are related to performance aboard ship. These variables are:

1. *Elementary school training,* as contrasted with non-school training. The number of records showing data on advanced service school success was too meager to permit an investigation of their relation to shipboard performance.
2. *Time lost between school and assignment to duty.* These data were secured only for the purpose of aiding in the interpretation of other data.
3. *Navy experience in rating and Navy experience not in rating.* An individual who has been engaged in radio work for a period of several months would be expected to be more proficient than an individual who had not had this amount of practice. Also, the fact that a man has been aboard ship for a period of time should be of some advantage when he becomes a "striker." The latter factor was, however, not included in the statistical analysis.
4. *Pay grade.* This variable may influence performance aboard the ship since the job assignments within a rating are often made in terms of the pay grade hierarchy. A signalman striker, for example, will rarely be allowed to send or receive flashing light messages on large ships regardless of his ability because these duties are usually the prerogative of the third class signalman.

5. *Effort put out by men on the job,* as estimated from comments of supervisors. The meaning and significance of these comments are not easy to assess. The comments may indicate merely that individuals with good work habits were liked or were wanted by their supervising petty officers, or that the men who were liked were also said to have good work habits.

6. *The length of time aboard a ship.* This variable is a possible prestige factor which may influence evaluational judgments, since a supervising petty officer may be tempted to rank higher the one of two equally competent individuals who has been aboard the ship longer.

Sources of Data Collected. The sources of the data included in this study were (1) the enlisted personnel Service Records which are filed aboard the ship with a duplicate copy in the Bureau of Naval Personnel, (2) the Enlisted Personnel Qualifications Card prepared by classification interviewers at recruit training stations and at advanced classification centers, and (3) the division officer's and chief petty officer's records. The data were collected either by the investigators themselves or by yeomen and classification interviewers not attached to the ship. This procedure resulted in a high degree of uniformity in the recording of the data, reduced clerical errors, and facilitated statistical processing.

The Enlisted Personnel Qualifications Cards were the only source of information concerning test scores and occupational experience data. Since all of the men had not been processed by the standardized classification procedures, some of the cards lacked scores on the Basic Test Battery, the civilian occupational codes as recorded were not always directly comparable, and the cards included notes and special test scores which were meaningful only to the classification unit preparing the original cards. Furthermore, Qualifications Cards prepared at advanced classification centers either contained no test scores or had only General Classification Test scores and, in a few cases, Arithmetic and Mechanical Aptitude Test scores. The failure to file duplicate copies of these Qualification Cards in the Bureau of Naval Personnel and the conception of these cards as non-official personnel records, separate from the Service Records, resulted in losses of data through careless handling and storage. The lack of Qualifications Cards reduced greatly the number of usable cases in the statistical analysis of the data.

The Criterion of Shipboard Performance

One of the most challenging problems encountered in this study was the development of an adequate criterion of shipboard performance. Cursory inspection led to the conclusion that no satisfactory ready-made criterion was available. Quarterly marks (in

proficiency in rating, seamanship, mechanical ability, ability as a leader of men, and conduct) were not sufficiently discriminative. Speed of advancement in rating was felt to be dependent upon too many irrelevant administrative factors to serve as a valid index. The paragraphs that follow will describe the development of a criterion of shipboard performance and the gathering of the criterion data.

PRELIMINARY PLANNING. In the planning stages of the project a number of decisions were made which defined the general character of the criterion to be developed.

First, it was decided that, in addition to seeking an overall measure of the quality of shipboard performance, a special effort should be made to secure a measure of a man's ability to do his job when he applied himself—his technical competence, as it was termed, irrespective of his personal qualities of leadership, dependability, etc. This decision was based on the fact that most of the objective indices used in selection and classification of enlisted personnel are designed primarily to predict technical competence rather than personal qualities, and that technical competence would therefore be the most useful criterion for evaluating the available prediction variables. In thus placing major emphasis upon technical competence, the assumption was made that technical competence is a significant factor in overall shipboard performance. Some evidence was obtained substantiating this assumption, although the matter was not submitted to rigorous test.

Second, it was decided not to administer objective achievement measures to men aboard ship. Administrative difficulties and differences between ships, as well as lack of time, precluded the possibility of developing and using a set of trade tests.

Third, the use of ratings by peers or associates was ruled out because it was believed that a true "peer" situation does not exist within the ratings, stratified as they are by pay grade from seaman second class up through the various petty officer categories.

Fourth, it was decided that the judgment of supervisors offered the most promising possibility for development of satisfactory criteria. Later experience bore out the initial hunch that, of the possible observers aboard ship, those best qualified to judge the men's effectiveness, particularly in the very important matter of having adequate opportunities for observation, are the supervising petty officers. Experience also indicated that these petty officers did not ordinarily have access to the data on the selection variables.

Within this general framework, tentative procedures were set up and progressively refined on the basis of a series of experimental try-outs, in order to answer two crucial questions concerning the criterion:

1. Can a satisfactory degree of agreement be obtained between two supervisors responding independently?

2. Can supervisors differentiate effectively between various aspects of performance aboard ship?

PRELIMINARY EXPERIMENTATION. In the course of five or six weeks of intensive experimentation aboard eleven ships returned from combat for repairs, the three following observations were made:

1. In general, rating scales proved relatively ineffective. Respondents apparently used such dissimilar standards that their ratings on 3- and 5-point scales showed sizeable differences in means and standard deviations, and a low degree of correlation between ratings by pairs of judges. This phase of the study was not exhaustive; in fact, late in the study a promising lead was found that might be refined to yield a satisfactory scale.

2. An adaptation of the "nominating" technique was tried, but it seemed to be ill-suited for this particular situation. Petty officers were asked to name the enlisted man whom they would most like and least like to take with them if they were transferred to another ship as the leading petty officer of their rating. They were then asked to state the reasons for their choices. In the time devoted to such efforts, it proved impossible to secure unambiguous categories. While continued effort might have produced appreciable improvement in the caliber of results obtained through this approach, the following observations, which led to its abandonment in this case, would have to be considered in any future project with similar respondents: (1) A sizeable proportion of the petty officers interviewed had great difficulty in expressing themselves—free response yielded practically nothing; confronted with check lists, they tended to accept or reject all suggestions with little apparent discrimination. (2) Even among the more verbally facile, two respondents would describe what appeared to be the same type of behavior but would choose quite different categories on a check list developed from their own responses.

3. Order-of-merit rankings appeared to be the most promising of the techniques utilized. Petty officers experienced little difficulty in ranking their men when a so-called "directed ranking" procedure was used. In this process, in order to control the set of the respondent toward his task, the interviewer retained the initiative throughout the ranking, presenting in random order the names of the men to be ranked, defining the attributes on which men were to be ranked, requiring explicit man-to-man comparisons, and periodically requesting the respondent to state why one man deserved to be ranked higher than another.

THE CRITERION ADOPTED. The criterion used in gathering the

data consisted of order-of-merit rankings in three areas: petty officer qualities, technical competence, and overall desirability. A high quality petty officer was defined as one "who obeys orders willingly, has initiative, takes responsibility, tries to do a good job, supervises men well, gives orders well." A poor quality petty officer was defined as one lacking these qualities. Technical competence was defined as "what a man knows about his job, what he can do when he really tries; not how he gives and takes orders, not his petty officer qualities." Specific definitions of "what a man knows about his job" were prepared for each of the six ratings. In the case of fire controlmen for example, the following four points were stressed: operation; checks and inspections; trouble shooting; and repairs. By thus placing emphasis upon specific aspects of performance it was possible to insure uniformity in the definition of technical competence from judge to judge and ship to ship.

Overall desirability was defined as follows: "If you were transferred to another ship from the one you are now on and were to be the senior petty officer, you would be interested in both the petty officer qualities of your men and their ability in rating. Considering both petty officer qualities and ability in rating, which one of these men would you most want to take with you if you were transferred to another ship? Which man would be of greatest value to you?" After the judge selected the best man he was asked to select the next best man and so on until all men in the group had been ranked.

In order to facilitate the data gathering process the definitions were written on cards so they could be read to the judge before he undertook to rank his men. Any questions which he might have were answered by the interviewer by providing further definitions or examples.

While the study was underway, checks were applied to the data obtained from eleven ships to determine the agreement between raters and the success with which supervising petty officers were able to discriminate between the three criteria: petty officer qualities, technical competence, and overall desirability. The rank-order correlation coefficient (*rho*) was used to determine these relationships. Checks on the agreement between judges in the technical competence rankings showed a median *rho* of .80. Those on petty officer qualities and overall desirability yielded a median *rho* of .60. In every rating the technical competence median *rho* was higher than the median *rho* for the petty officer qualities and overall desirability rankings.

Checks on the discrimination among the three criteria showed a median rank-order correlation coefficient between petty officer qualities and technical competence of .80, indicating at best only a slight

degree of differentiation. The rank-order correlation coefficients for petty officer qualities with overall desirability ranged from .83 to .91, and those for technical competence with overall desirability from .80 to .95. These values tend to be as high as the reliability of the separate rankings would permit. There was some discrimination for certain groups, particularly radiomen, and for that reason rankings on all three criteria were gathered throughout the study. In the final analysis of the data, however, the criterion used was the technical competence rankings adjusted by the procedure to be described in a later section.

Procedures Followed in Gathering the Data

As stated earlier, the data for this study were gathered aboard ships which had returned to the continental United States from forward areas for overhaul or repairs. The actual steps in the procedure were as follows:

1. The project was explained to the commanding officer or executive officer or both and arrangements made for the time and place for meeting the supervising petty officers in each rating. Usually the executive officer in consultation with the division officers selected the two judges who were best qualified to rank the men in the rating and arranged the schedule for the interviewers.

2. The names of the men in the groups were written on specially prepared 5″ x 6″ cards. These cards provided spaces on one side for recording and coding the following data: test scores, age, occupational data, quality classification, classification interviewers' recommendations, amount of civilian education, Navy training data, advancement in rating data, shipboard experience, and name of present ship. On the reverse side, spaces were provided for recording the ranks assigned by supervising petty officers and any special comments concerning work habits, personality etc. At the time the cards were used in the ranking process, only the names of the ratees appeared on them. A precaution which had to be observed was to record only the names of men known to both supervising petty officers. This usually resulted in the elimination of a few men in the group.

3. The nature of the project was explained to each judge individually. Emphasis was placed on the confidential nature of the project and assurance was given that the data would be used for research purposes only. This pledge was reenforced by a letter from the Chief of Naval Personnel which stated that no administrative use would ever be made of the data; this letter was shown to each judge. With very few exceptions, the judges seemed to accept these statements and entered into the project in a cooperative and enthusiastic manner.

4. The nature of "petty officer qualities" was explained to the judge, and the card defining this criterion was placed before him. A data card

was selected at random and placed on the table. A second card was picked at random and the judge asked "Is this man better or not as good as the first man?" If the reply was "better," the card was placed above that of the first man; if the reply was "not as good," the card was placed below. This procedure was continued until all cards had been placed on the table in rank order. The judge was then asked to examine the rank order carefully and make any changes which he felt should be made. In order to assist the judge in the checking process the interviewer asked this question about a number of men, "Exactly why is A better than B?" The interviewer then recorded these ranks on the backs of the data cards.

5. A similar procedure was followed in getting the technical competence rankings (the cards being shuffled before beginning the new rankings). As described earlier a card was prepared for each rating defining the nature of technical competence in that rating. During the ranking process, the judge was periodically reminded of the definition of technical competence and warned not to confuse petty officer qualities with technical competence.

6. The rankings in overall desirability were obtained by substantially the same procedure as that used in (4) and (5) above. First the judge selected the man whom he would most like to take with him if transferred to another ship, then the next best man, and so on until all the men had been placed in rank order. During the ranking process the judge was warned not to judge the men on a personal basis but on their overall value to the new ship.

7. During the later stages of the study, ratings were obtained for each man in the group on a twenty point scale based on the Navy pay grade system—striker, third class, second class, first class. Each of these categories was then divided into five steps. The following instructions were given:

> Forget the pay grade each man actually holds and consider only his ability in terms of all men in rating in the Navy today whom you know. Select (the scale was placed before the judge) the number that best fits your idea of each man's ability. For example, a third class man may actually be better than most second class men in the Navy today; if so, he would be given a 14 or 15 on this scale; on the other hand a second class man may only be as good as the average striker and would be given a 3. The highest rating which you can give is 20 and 1 is the lowest.

Although these ratings were not used in the final criterion, they constitute a promising lead as to a type of rating scale which may prove useful in evaluating the performance of enlisted personnel.

8. The judge was instructed not to discuss his rankings with other personnel aboard ship and was given assurance that the data would be used for research purpose only.

9. Steps 3-8 inclusive were followed with the second judge.

10. The background and experience data were copied on the data cards (described in 2 above) from the Service Records and the Enlisted Personnel Qualifications Cards.

11. The ranks were converted to scale values by Hull's method (after correction for unreliability of judgments as described in the next section) [1], the criterion values for the two judges averaged, and the data on the cards coded in order to transfer the data to IBM punched cards. The main statistical analysis was made by use of the IBM cards.

Refinement of the Criterion Data

The technical competence ratings which were accepted as the criterion measure for the statistical analyses in this study were contaminated by two extraneous factors: (1) unreliability of judgments and (2) differences in Navy experience of the ratees. It seemed that more valid data on the relationship between background factors and quality of performance aboard ship could be obtained by reducing the effect of those factors. Therefore steps were taken to make the necessary statistical adjustments in the criterion measures.

In all, 120 rating groups of men with a total of 1,977 cases were used in the study. The groups ranged in size from 6 to 33 cases, with a median of 17. Rank-order correlation coefficients indicating the amount of agreement between pairs of respondents on technical competence were computed for all groups. These *rho*'s ranged from .08 to .98 for the 120 groups, with median *rho*'s for the six ratings ranging from .78 to .91.

It was decided that only rating groups for which the judges were in substantial agreement would be used in the analysis. In a number of the groups the disagreement between judges was confined to a relatively small number of cases. A procedure was therefore set up to evaluate the acceptability of the data as follows: First, the rank-order correlation coefficient for all cases in a rating group was determined and a distribution of rank-order differences was made. Single cases were then eliminated if the following threefold criterion were met:

1. if the rank-order difference was greater than N/3,
2. if the distribution of differences showed marked positive skewness, together with a discontinuity at a value less than N/3, and
3. if the comments of the judges indicated definite disagreement or revealed reasons for anticipating disagreement in that particular case.

After such cases were eliminated, the rank order correlation coefficient was recomputed for each group and only those groups were retained for which the *rho* was greater than .70. Nine groups of the

[1] Hull, Clark L. *Aptitude Testing*, Appendix 1. World Book Co., Yonkers, N. Y., 1928.

original 120 were eliminated in this way and 44 single cases were eliminated from 26 of the remaining 111 groups. After the elimination of these groups and single cases, the median *rho*'s indicating agreement on technical competence ranged from .84 to .91 for the six different ratings.

For any measure of performance, positive correlation between amount of experience on the job and the criterion of success would be expected because of the increase in opportunities for learning associated with longer service and more responsibility. Median rank-order correlation coefficients between number of months of experience in rating and technical competence values varied from .48 for the fire controlman groups to .69 for the radioman groups. Since the relative degree of success of the men, independent of the length of time they had been in rating, was the criterion desired, an adjusted criterion score was secured by partialling out experience from the technical competence values.

The general procedure was to estimate the correlation coefficient between months of experience in rating and the technical competence scale values. The regression coefficient corresponding to this correlation coefficient was then used to adjust each person's technical competence scale value in such a way that his relative standing on the adjusted value scale would be independent of the amount of his experience.

Through applying the equation for the regression of technical competence values on experience, predicted values were obtained. The distribution of these values represented the variations in technical competence scale values that could be attributed to the variations in experience in rating. When these adjusted values, taken as deviations from the mean, were subtracted from the corresponding obtained technical competence scale values, also expressed as deviations from the mean, the result was the desired adjusted technical competence scale values.[2]

In the solution of this particular problem, the adjustment was made separately for each ship group in each rating. The correlation coefficients used in making the adjustment were also estimated separately for each such group.

It was not considered appropriate to use the average values for all ship groups for a single rating because tests of homogeneity revealed greater than chance differences between ship groups both with respect to the variance of months of experience in rating and with respect to the correlation coefficients between experience in rating and technical competence values.

[2] The use of this technique assumes a zero correlation between test scores and experience. This assumption was approximately fulfilled in these data.

The correlation coefficients between experience and technical competence that were used in making the adjustment were themselves adjusted values. The adjustment was made as follows. The correlation coefficient between (1) standard deviation of months of experience and (2) the correlation coefficients of months of experience with technical competence were determined for all ships in each rating group. The correlation for experience with technical competence was then adjusted for each ship so that it would fall on the regression line at the point corresponding to the actual experience standard deviation of the ship.

Adjusted technical competence values, consisting of the residuals with the influence of experience held constant, constitute the final criterion developed.

Evaluation of Criteria Adopted. The criterion finally adopted possessed both strengths and weaknesses. The criterion stands up well in the following respects: First, it has what may be termed "face validity". The judgments were requested on the actual performances for which the men had been selected and trained, and judgment of petty officers is used regularly in the Navy in evaluating personnel. The particular petty officers interviewed in this study were in a position to observe the subjects and had good reason to do so aside from the study itself. Second, the criterion was reasonably stable—two or more qualified observers agreed to an extent roughly comparable to the estimated reliability of the selection variables. Third, the criterion is reasonably uncontaminated with regard to the selection variables—there is no reason to suspect that the judges were familiar, except in a very general way, with the classification and training data on their men or that they were interested in the specific outcomes of the study.

On the other hand, certain definite limitations should be recognized: First, evidence is lacking as to the degree to which petty officers' judgments on achievement correspond to individuals' actual performance on objective measures of such achievement. Second, the data are not comparable from ship to ship, necessitating a completely "within ships" analysis with very small numbers. It should be noted, however, that this difficulty is not a function of the ranking method *per se* but simply reflects the absence of any systematic provision in the Navy for supervisors to observe the work of specific men on more than one ship over a given period of time. Third, both means and standard deviations of the criterion are arbitrarily determined in advance when the ranking method is used with a limited number of judges; this limitation forestalls any statistical analysis involving the variance of the criterion, thereby increasing greatly the computational labor required in the analysis and com-

plicating the problem of estimating the reliability of the statistics. Fourth, no differentiation was obtained between petty officer qualities and technical competence. Evidently the technical competence rating obtained was in reality a sort of overall rating including personal qualities as well as technical proficiency.

Analysis of Relation of Criterion to Predictive Variables

The results of this study can be grouped into two major categories. First are the statistical data which describe the population used in this study; second are the data showing the relationship between background factors and criterion. Before presenting the results, a brief description of the statistical procedures used in analyzing the data will be given.

STATISTICAL PROCEDURES. The statistical procedures used in analyzing the relationships between background factors and the adjusted criterion values (variance due to experience eliminated) were chiefly correlational analyses and *chi*-square tests. Each of the procedures is listed and described briefly below.

1. Rank order correlation coefficients (*rho*'s) were used in the preliminary analyses of the criterion, while product-moment correlation coefficients were computed after the criterion values were expressed in average linear scale values. These product-moment coefficients, termed within-group correlations, were computed separately for each ship rating-group.

2. A single value representative of the within-group correlations was determined for each rating from the weighted average correlations found between two variables, using the Z transformation technique. The separate Z-values were weighted by an appropriate number (N-3) in determining the average Z-value. The bias in the mean Z-value is not great enough to warrant the use of correction procedures in this study.[3]

3. The hypothesis that the samples from which the correlations between two variables were determined could have come from equally correlated populations was tested in each case in which an average within-group correlation coefficient was determined. This hypothesis was tested by determining the variance of the weighted squared deviations of the Z's from the mean of the Z-values for the several groups being combined. The test of the hypothesis is based upon the *chi*-square distribution of the quantity $\Sigma(N\text{-}3)Z^2 - \frac{(\Sigma[N\text{-}3]Z)^2}{\Sigma(N\text{-}3)}$ summed over the set of samples.[4] This test and the averaging of Z-values should be used when the number of samples is small compared to the average size of the

[3] Rider, Paul R. *An Introduction to Modern Statistical Methods*, p. 106. John Wiley & Sons, Inc., New York, 1939.

[4] Rider, *op. cit.*, p. 107.

sample, a condition which was not generally obtained in this study. The rigor of this test in indicating whether the null hypothesis should be accepted or rejected is that of the *chi*-square test and may not be sufficiently sensitive to obtained differences.

4. Product-moment correlation coefficients based on all cases combined in a single group were also determined between a selected number of variables and the criterion values for the particular ratings. Comparisons of the average within-group validity coefficients with these combined data coefficients indicate the extent to which variation in the means of the groups on the prediction variables affects the correlation coefficients.

5. The *chi*-square test, assuming the null hypothesis, was used to investigate the relationship between the criterion values and such categorized variables as interviewer's quality classification code, elementary school recommendation, civilian occupation, and estimated level of effort. The relationship of the latter to the criterion was also tested by means of the coefficient of contingency.

6. The relationship between the type of training and the adjusted criterion values was analyzed by two procedures for each rating: (1) by the biserial correlations between the type of training (school or non-school) and the criterion values, and (2) by a co-variance analysis of the differences in mean criterion scores of school and non-school trained groups. For this analysis, the two groups in the rating were equated in terms of mean score on the one test of the Basic Test Battery which correlated highest with the criterion for that rating. The ratio of the numbers of cases in each of the two groups (school and non-school) for each ship sample was held constant by selecting random samples for each group through the use of tables of random numbers.

CHARACTERISTICS OF THE POPULATION. The distribution of the sample population by ship type and rating is shown in Table 1-xx. The number of cases listed in this table is that which remained after all unusable cases had been deleted by the procedures described in the preceding section. The total number of cases included in the final analysis is 1,868 from 27 ships. The cases are distributed fairly well among the three types of ships and among the six ratings.

The median numbers of cases per group in each rating are as follows: 17 for radioman groups, 15 for signalman, 16 for radarman, 15 for fire controlman, 16 for machinist's mate and 18 for gunner's mate groups. Unfortunately the number of usable cases in any particular statistical analysis varied considerably owing to incomplete records.

Table 2-xx shows for the sample population the average score in three selection tests, average age, and average number of years of civilian education by ship type and rating. The highest average scores on the Basic Test Battery are those of the fire controlmen;

the lowest scores in these tests are made by the gunner's mates. Radiomen, radarmen, and fire controlmen are very similar with respect to amount of civilian education; gunner's mates have the lowest average number of years of education. With respect to age the signalmen and radiomen are the youngest and radarmen and

TABLE 2-XX. Mean score on three selection tests, mean age, and mean number of years of civilian education by ship type and by rating for the sample population [1]

Rating	Ship Type	Mean Test Score			Mean Age in Years	Mean Years Civilian Education
		GCT	ARI	MKM		
RM	Destroyer	52.4	51.0	50.4	21.9	11.0
	Carrier	53.4	49.2	49.2	21.5	11.1
	Cruiser	54.0	46.6	45.9	21.7	11.2
	Total	53.3	48.3	48.3	21.7	11.1
SM	Destroyer	48.0	41.4	47.2	22.9	10.7
	Carrier	49.9	46.8	46.6	21.5	10.5
	Cruiser	52.0	47.6	46.4	21.2	10.7
	Total	50.9	47.1	46.6	21.5	10.6
RdM	Destroyer	54.4	50.6	48.0	26.0	11.4
	Carrier	52.9	50.6	52.0	25.6	11.5
	Cruiser	54.1	52.1	50.9	24.5	11.3
	Total	53.7	51.1	50.7	25.4	11.4
FC	Destroyer	58.8	53.2	56.5	23.4	11.4
	Carrier			No Cases		
	Cruiser	54.4	53.3	53.6	23.7	11.5
	Total	56.2	53.3	54.2	23.6	11.5
MM	Destroyer	50.3	47.9	53.5	24.1	10.3
	Carrier	50.4	48.4	54.7	24.5	10.6
	Cruiser	51.4	48.6	56.8	23.7	10.7
	Total	50.6	48.4	55.1	24.2	10.5
GM	Destroyer	51.2	46.3	53.6	23.6	10.2
	Carrier			No Cases		
	Cruiser	44.2	42.0	45.8	23.5	9.3
	Total	46.6	43.4	48.8	23.6	9.7

[1] The mean values in this table are based upon differing numbers of cases. All available data for these variables were included in the computations.

machinist's mates are the oldest. In general it appears that this population is fairly representative of rated men in the Navy.

RELATIONSHIP OF EXPERIENCE TO THE CRITERION. In the discussion of the development of the criterion it was stated that there was a substantial relationship between experience and the average technical competence scale values and that these scale values had to be

adjusted for differences in amount of experience. The data concerning average amount of experience and variability in experience by rating and ship type are shown in Table 3-xx. Aboard destroyers, the machinist's mates are most experienced, while aboard carriers and cruisers and in the combined group, radiomen are most experienced.

TABLE 3-xx. Months of experience in rating by type of ship for six ratings: means and standard deviations of distribution

	Destroyer			Carrier			Cruiser			All Ship Types		
Rating	N	M	σ	N	M	σ	N	M	σ	N	M	σ
RM	72	19.4	8.28	172	20.8	9.28	122	23.4	8.38	366	21.4	8.93
SM	12	19.5	8.73	104	18.6	7.46	82	19.0	8.91	198	18.8	8.17
RdM	77	14.6	4.29	194	16.6	5.01	94	16.5	7.35	365	16.1	5.64
FC	75	18.9	7.76	No Cases			144	19.7	7.33	219	19.4	7.49
MM	87	22.6	9.59	212	16.8	7.51	128	19.5	9.17	427	18.8	8.77
GM	137	21.5	9.04	No Cases			156	17.0	8.90	293	19.1	9.24

The coefficients of correlation between amount of experience and the criterion are shown in Table 4-xx. It should be noted that these are average within-group coefficients; in the case of radiomen, for example, the correlation is the average of the twenty-one which were computed for this group. All of the correlations are quite high and indicate that there was a strong tendency for the most experienced men to get the highest ratings, thus substantiating the contention that the criterion values should be adjusted for experience.

TABLE 4-xx. Average within-group correlation coefficients between months of experience in rating and unadjusted technical competence average scale values for six ratings

Rating	Number of Ships	Number of Cases	P-value[1] (Homogeneity)	Average Correlation Coefficient
RM	21	366	50-70%	.69
SM	14	198	<1%	.69
RdM	20	365	<1%	.52
FC	13	219	2-5%	.48
MM	26	427	1-2%	.65
GM	17	293	<1%	.55

[1] The values in this column were obtained from a *chi*-square test of the hypothesis that all ship groups within a rating were random samples from equally correlated populations. The P-value indicates the probability that a dispersion of correlation coefficients as large as was obtained would occur among such random samples by chance alone.

RELATIONSHIP OF BASIC TEST BATTERY SCORES TO THE CRITERION. Average within-group correlations between the six tests of the Basic Test Battery and the criterion values were determined separately for each rating. These data are presented in Table 5-xx. The significance of each Z-value was evaluated in terms of its standard error; those values lying between the 1 per cent and 5 per cent levels of significance are in italics, those beyond the 1 per cent level are in bold face. This table indicates that the six tests of the Basic Test Battery tend to have significant positive correlations with the criterion. The one negative correlation, between the Reading Test and the criterion for gunner's mates, was not significantly different from zero.

The pattern of correlation coefficients in Table 5-xx is approximately what one would expect. For radiomen, signalmen, and radarmen the verbal and arithmetical tests show the highest correlations with the criterion. In the case of gunner's mates, machinist's mates, and fire controlmen, the highest correlations are those between one or more of the mechanical tests and the criterion. In every rating there is at least one correlation over .30; in the case of radarmen one test (Arithmetical Reasoning Test) correlates .44 with the criterion. Although these correlations are not as high as those found between Basic Test Battery scores and success in elementary and advanced enlisted schools, they follow the same general pattern. Evidently the test scores can be used with some degree of success to predict quality of performance aboard ship.

The correlation coefficients given in Table 5-xx are average correlation coefficients; that is they are weighted averages of the coefficients obtained for the separate groups in each rate. The correlation of the General Classification Test with the criterion in the case of radiomen, for example, is the average of the thirteen coefficients computed for this rating. This procedure was followed because the groups from the different ships differed significantly in average General Classification Test scores; hence it was felt that the thirteen groups should not be combined into one population and a single correlation coefficient computed between General Classification Test scores and the criterion.

In Table 6-xx are shown the correlation coefficients between scores on three tests of the Basic Test Battery and the criterion, computed by the "within-group" and "combined" techniques. Although the differences are not large, it can be seen that in general the correlation coefficients computed by the within-group technique are highest. Evidently combining the different groups into one population (by rating) results in general in an attenuation of the correlation coefficients.

TABLE 5-XX. Average within-group correlation coefficients between Basic Test Battery scores and adjusted criterion values for six ratings

Rating	Tests	Number of Ship Groups	Total Number of Cases	Average Correlation Coefficient with Criterion	Average[1] Weighted Z-Value	Standard Error of Average Z-Value
RM	GCT	13	175	.38	**.403**	.09
	READ	12	126	.33	**.338**	.11
	ARI	12	143	.32	**.335**	.10
	MAT	10	118	.16	.158	.11
	MKM	11	125	.26	**.270**	.10
	MKE	11	125	.16	*.165*	.10
SM	GCT	7	97	.31	**.317**	.12
	READ	3	33	.02	.026	.20
	ARI	5	69	.29	*.301*	.14
	MAT	4	57	.31	*.325*	.15
	MKM	5	59	.28	*.292*	.15
	MKE	5	59	.10	.108	.15
RdM	GCT	17	261	.31	**.320**	.07
	READ	16	228	.26	*.271*	.07
	ARI	16	234	.44	**.478**	.07
	MAT	16	236	.26	**.268**	.07
	MKM	16	232	.15	*.151*	.07
	MKE	16	232	.26	**.260**	.07
FC	GCT	8	154	.23	**.238**	.09
	READ	6	99	.20	*.208*	.11
	ARI	6	122	.24	**.253**	.10
	MAT	6	115	.33	**.345**	.10
	MKM	6	100	.19	*.193*	.11
	MKE	6	100	.24	*.246*	.11
MM	GCT	19	236	.18	**.187**	.08
	READ	14	139	.03	.027	.10
	ARI	14	159	.25	**.251**	.09
	MAT	15	158	.13	.130	.09
	MKM	15	153	.36	**.375**	.10
	MKE	15	153	.31	**.323**	.10
GM	GCT	11	166	.15	.150	.09
	READ	5	52	—.11	—.112	.16
	ARI	9	121	.27	*.275*	.10
	MAT	6	83	.35	**.369**	.12
	MKM	6	74	.36	**.375**	.13
	MKE	6	74	.29	*.300*	.13

[1] The significance of these average Z-values was evaluated in terms of the ratio of the average to its standard error, σ_z; values lying between the 1% and 5% levels are in italics, those beyond the 1% level are in bold face.

The intercorrelations between three tests of the Basic Test Battery, age, and years of civilian education are shown in Table 7-xx. The mean and standard deviation for each variable by rating group are also presented. In general the intercorrelations between the three Basic Test Battery scores are somewhat lower than those reported in Chapter VI, probably because these groups have been selected in part on the basis of test scores and show somewhat less variability than a recruit population.

Relationship Between Age and the Criterion. As shown in Table 8-xx there is a low but significant correlation between age and the criterion for the radioman, radar operator, machinist's mate, and gunner's mate groups. The mean age and standard deviation for

Table 6-xx. Comparison of the correlation coefficients between scores on three tests of the Basic Test Battery and the criterion for six ratings, computed by the combined group and average within-group procedures

Rating	Method	Number of Cases	Test		
			GCT	ARI	MKM
RM	combined groups	156	.33	.18	.09
	within-group	varied	.38	.32	.26
SM	combined groups	96	.22	.24	.23
	within-group	varied	.31	.29	.28
RdM	combined groups	248	.29	.39	.12
	within-group	varied	.31	.44	.15
FC	combined groups	117	.25	.25	.18
	within-group	varied	.23	.24	.19
MM	combined groups	206	.16	.21	.31
	within-group	varied	.18	.25	.36
GM	combined groups	124	.17	.16	.29
	within-group	varied	.15	.27	.36

each rating are shown in Table 7-xx. Although age requirements are frequently specified for various billets, there is little evidence in this table to indicate that this factor should be given serious consideration in the assignment of men, at least not within the age range of the men included in this study.

Relationship Between Amount of Education and the Criterion. The means and standard deviations of years of civilian education of men in different ratings are shown in Table 7-xx. The relationship between years of education and the criterion values is shown in Table 9-xx. These average within-group correlations range from .01 for radiomen to .27 for signalmen. Whether these coefficients represent anything different from the common abilities or

capabilities represented by the Basic Test Battery scores cannot be established from the data presented in this chapter. Data in Chapter VI and in Table 7-xx show, however, that there is substantial correlation between Basic Test Battery scores and amount of education. It does not seem likely, therefore, that amount of education would

TABLE 7-xx. Intercorrelations, means, and standard deviations of selected variables computed for all cases in each rating for which complete data on these variables were available

Variable	Rating Group	Number of Cases	Intercorrelations					M	σ
			GCT	ARI	MKM	Age in Years	Years of Civilian Education		
GCT	RM	156		.60	.29	.23	.26	52.2	7.93
	SM	96		.56	.30	.10	.64	50.5	8.32
	RdM	248		.68	.35	.34	.38	53.5	8.47
	FC	117		.72	.48	.26	.49	56.5	9.28
	MM	206		.68	.58	.06	.51	49.6	8.89
	GM	124		.61	.60	.10	.55	44.9	9.81
ARI	RM	156	.60		.36	.21	.18	48.4	9.09
	SM	96	.56		.35	.14	.62	47.1	10.03
	RdM	248	.68		.31	.25	.33	51.0	10.69
	FC	117	.72		.48	.28	.41	54.1	12.16
	MM	206	.68		.44	—.03	.39	48.6	10.12
	GM	124	.61		.49	.05	.34	43.9	8.74
MKM	RM	156	.29	.36		.27	.09	48.2	8.48
	SM	96	.30	.35		.36	.17	46.4	9.55
	RdM	248	.35	.31		.29	.09	50.8	10.25
	FC	117	.48	.48		.41	.30	54.3	9.98
	MM	206	.58	.44		.25	.42	55.1	9.72
	GM	124	.60	.49		.24	.42	49.0	10.27
Age in Years	RM	156	.23	.21	.27		.22	20.6	2.31
	SM	96	.10	.14	.36		.14	21.1	3.31
	RdM	248	.34	.25	.29		.08	25.8	6.27
	FC	117	.26	.28	.41		.13	23.7	4.87
	MM	206	.06	—.03	.25		—.21	24.7	5.46
	GM	124	.10	.05	.24		—.08	23.2	4.36
Civilian Education in Years	RM	156	.26	.18	.09	.22		11.2	1.21
	SM	96	.64	.62	.17	.14		10.8	1.39
	RdM	248	.38	.33	.09	.08		11.5	1.51
	FC	117	.49	.41	.30	.13		11.5	1.30
	MM	206	.51	.39	.42	—.21		10.6	1.71
	GM	124	.55	.34	.42	—.08		9.4	2.13

be a very useful selection device if an individual's Basic Test Battery scores are available.

RELATIONSHIP BETWEEN NAVY TRAINING AND THE CRITERION. Navy training schools exist for the purpose of preparing men for their shipboard billets. It is of interest, therefore, to know how

TABLE 8-XX. Average within-group correlation coefficients between age in years and adjusted criterion values for six ratings

Rating	Number of Ships	Total Number of Cases	Average Z[1]	σ_z	Average r with Age in Years
RM	21	366	*.148*	.06	.15
SM	12	186	.127	.08	.12
RdM	20	365	**.190**	.06	.19
FC	13	218	.045	.08	.04
MM	25	410	*.142*	.06	.14
GM	17	293	**.164**	.06	.16

[1] The significance of these average Z-values was evaluated in terms of the ratio of the average to its standard error, σ_z; values lying between the 1% and 5% levels are in italics; those beyond the 1% level are in bold face.

school trained men compare with non-school trained personnel in performance aboard ship. As explained in Chapter V, men who do not go to school can acquire the necessary training to qualify for petty officer status through the in-service training program. It is conceivable that the latter type of training may be as good as or even better than training given ashore. Data were gathered in this study to throw some light on this question.

A complicating factor in any study of the relationship between school training and success aboard ship is the fact that school populations average higher in ability test scores—having been selected on that basis. Table 10-xx shows that the school trained personnel average higher in all tests of the Basic Test Battery than do non-

TABLE 9-XX. Average within-group correlation coefficients between years of civilian education and adjusted criterion values for six ratings

Rating	Number of Ships	Total Number of Cases	Average Z[1]	σ_z	Average r with Years of Civilian Education
RM	21	361	.009	.06	.01
SM	12	186	**.282**	.08	.27
RdM	20	361	*.121*	.06	.12
FC	13	215	**.232**	.08	.22
MM	26	423	**.196**	.05	.19
GM	17	291	**.202**	.06	.20

[1] The significance of these average Z-values was evaluated in terms of the ratio of the average to its standard error, σ_z; values lying between the 1% and 5% levels are in italics; those beyond the 1% level are in bold face.

TABLE 10-xx. Means and standard deviations of scores on three Basic Test Battery tests for school, non-school, and combined groups for six ratings

		Test								
		GCT			ARI			MKM		
Rating		N	M	σ	N	M	σ	N	M	σ
	School	145	55.7	7.60	116	49.8	9.10	105	48.7	8.50
RM	N. Sch.	62	47.9	6.60	57	45.4	8.40	56	47.7	8.30
	Total	207	53.3	7.30	173	48.3	8.87	161	48.3	8.43
	School	104	58.6	7.92	85	56.5	10.52	73	57.7	11.24
FC	N. Sch.	71	52.6	9.12	58	48.5	12.20	48	49.0	8.44
	Total	175	56.2	8.92	143	53.3	11.92	121	54.3	9.88
	School	192	56.0	7.52	180	53.6	9.68	180	52.7	10.12
RdM	N. Sch.	85	48.3	8.16	73	44.9	10.12	72	45.8	9.08
	Total	277	53.7	8.52	253	51.1	10.60	252	50.7	10.32
	School	57	52.0	7.72	49	47.9	10.00	43	47.3	9.88
SM	N. Sch.	69	50.1	8.48	62	46.3	9.32	58	46.1	9.28
	Total	126	50.9	8.20	111	47.1	9.64	101	46.6	9.56
	School	124	54.8	7.70	93	52.2	10.10	89	57.8	8.90
MM	N. Sch.	143	47.0	8.30	120	45.5	9.00	114	53.0	9.90
	Total	267	50.6	8.02	213	48.4	9.48	203	55.1	9.46
	School	57	55.6	6.24	49	48.8	8.44	42	55.8	8.36
GM	N. Sch.	140	42.9	8.44	106	40.9	7.84	85	45.3	9.68
	Total	197	46.6	9.76	155	43.4	8.80	127	48.8	10.48

school trained personnel. Differences are most significant in the case of gunner's mates and least significant in the case of signalmen.

The relationship between Navy training and success aboard ship was studied by three different methods. In Table 11-xx are shown the mean criterion values for the school trained, non-school trained, and combined groups. The school trained groups average higher than non-school trained groups in every rating; but as pointed out in the preceding paragraph, this may simply be due to the fact that

TABLE 11-xx. Means and standard deviations of adjusted criterion measures for school, non-school, and combined groups for six ratings

	School Trained Men			Non-School Trained Men			Combined Group		
Rating	N	M	σ	N	M	σ	N	M	σ
RM	254	52.5	13.44	111	45.0	12.72	365	50.3	13.62
SM	86	51.6	13.74	106	48.8	14.34	192	50.1	14.22
RdM	249	52.5	15.54	113	45.4	14.04	362	50.3	15.48
FC	131	50.2	15.36	80	48.8	16.86	211	49.7	15.90
MM	193	51.8	14.64	220	48.5	13.56	413	50.0	14.16
GM	85	54.0	15.72	199	48.0	14.76	284	49.8	15.36

school trained personnel are of higher quality as measured by ability tests. In Table 12-xx are shown the biserial correlations showing the relationship between attendance or non-attendance at school and success aboard ship. Actually Tables 11-xx and 12-xx present substantially the same facts through different media—the first through difference in means, the second through correlational data. The biserial correlation coefficients as well as the mean criterion values are influenced by the fact that school trained personnel have higher ability test scores.

In order to hold constant the factor of ability, a covariance test was applied to the data, equating the groups on the basis of the Basic Test Battery score correlating highest with the criterion. Application of the F-test (Table 13-xx) shows that the only difference

TABLE 12-xx. Biserial correlation coefficients between school vs. non-school training and adjusted criterion values for six ratings

Rating	Number of Cases: School Trained[1]	Number of Cases: Non-School Trained	r_{bis}
RM	254	111	.34
SM	86	106	.12
RdM	249	113	.28
FC	131	80	.05
MM	193	220	.14
GM	85	199	.23

[1] School trained men attending any school other than one for the rating at which they were working were omitted.

between school and non-school trained men significant at the 1 per cent level is that in the case of radiomen. The difference for radarmen is significant at the 5 per cent level but not at the 1 per cent level. The smallest F-ratio occurred in the case of machinist's mates.

The relationship between success in elementary school and success aboard ship is shown in Table 14-xx. Elementary school success is measured in two ways—percentage grade earned and rank in class. Except in the case of signalmen, the correlation coefficients between rank in class and success aboard ship are higher than those between school grades and the criterion. This finding is in accordance with that usually found in civilian education. The relationships between school success and success aboard ship are substantial and indicate that in general the men who do well in school also do well aboard ship.

TABLE 13-XX. Significance of differences in mean adjusted criterion values for school and non-school trained groups, equated on mean scores of test of Basic Test Battery correlating highest with the criterion

Rating[1]	Test Correlating Highest with Criterion	Number of Ship Groups	Number of Cases		F-Ratio	Degrees of Freedom	P-Values
			School	Non-School			
RM	GCT	11	72	37	11.61	1 and 9	P<1%
SM	GCT	11	35	58	3.17	1 and 9	P>5%
RdM	ARI	15	126	27	6.64	1 and 13	1%<P<5%
MM	MKM	15	70	39	1.15	1 and 13	P>5%

[1] No covariance test could be made on fire controlmen and gunner's mates because of the small numbers of cases available.

RELATION BETWEEN EFFORT AND THE CRITERION. After the judges had completed the ranking of their men on petty officer qualities, technical competence, and overall desirability, they were asked to comment on the work habits and personal characteristics of the upper and lower 20 per cent. These comments were coded as follows: (1) no comment on effort, (2) little effort expended, (3) below average in effort, (4) average effort expended, (5) above average as a worker, (6) hard worker, (7) disagreement between judges on the work habits of the man. In a study of 308 cases it was found that codes assigned by coders working independently agreed in 84 per cent of the cases.

TABLE 14-XX. Average within-group correlation coefficients between two measures of school success and adjusted criterion values for six ratings

Rating	School Success Measure	Number of Ships	Total Number of Cases	Average r	Average Z	σ_z
RM	Grade	14	191	.40	.425	.08
	Rank in class	11	143	.51	.562	.10
SM	Grade	2	24	.52	.572	.24
	Rank in class	2	20	.34	.352	.27
RdM	Grade	15	199	.30	.312	.08
	Rank in class	13	184	.39	.410	.08
FC	Grade	8	95	.44	.470	.12
	Rank in class	7	83	.48	.528	.13
MM	Grade	10	94	.25	.251	.12
	Rank in class	9	83	.45	.482	.13
GM	Grade	5	49	.22	.225	.17
	Rank in class	2	20	.29	.295	.27

The relationship between effort (as coded) and the criterion (trichotomized at the one half sigma point above and below the mean) was determined by means of the coefficient of contingency. In the school trained groups, the contingency coefficients for the different ratings ranged from .28 to .51 with a median of .38; in the non-school trained groups, the coefficients for the different ratings ranged from .19 to .38 with a median value of .29. The coefficients in the combined groups ranged from .36 to .49 for the different ratings with a median value of .43.

The meaning of the relationships between effort and the criterion are difficult to interpret. Although the relationship between the three criterion measures and the estimates of effort were not computed, there are strong indications that the effort ratings were actually only different facets of the overall evaluations. In other words if a supervising petty officer thought well of a man he was likely to say that the man was a hard worker; if he didn't like the man he was likely to say that the man was a poor worker.

Other Factors. The relationships of the following other factors to the criterion were also studied: civilian occupation, quality classifications and recommendations made by classification interviewers, time lost between school and assignment to duty, Navy experience not in rating, and length of time aboard a particular ship. Since no significant relationships were found and since some of the data on the factors (notably information about civilian occupation) were of questionable quality, the results are not presented here.

Summary

The findings of this study can conveniently be grouped into two major categories: those pertaining to the criterion and those pertaining to the relationship between background factors and shipboard performance. Since this was the first systematic study undertaken by the Bureau of Naval Personnel of the relationship between classification and training data and shipboard performance, it was necessary to do considerable experimental work on the development of a suitable criterion. In many respects the major contribution of the study lies in the work which was done on the development of the criterion measures and in "purifying" them for use in statistical analysis.

Development of a Criterion. The high degree of agreement between supervising petty officers is the most significant feature of the criterion finally adopted. The relationships, expressed as rank-order correlation coefficients, are considerably higher than those generally reported in the literature for independent rankings of indi-

viduals based on their performance on the job. While it is true that the supervisors had probably discussed the capabilities of the men in their groups, there is no evidence that they had ever actually expressed these judgments as ranks. Certainly there was no prior opportunity for the supervisors to compare judgments pertaining to the evaluations requested of them in this study.

The most disappointing aspect of the research on the development of the criterion was the inability of the supervisors to differentiate between qualities of personality (petty officer qualities) and technical competence. Every effort was made during the ranking process to emphasize the differences between these two variables. While some differentiation was apparently obtained in the radioman groups, the relationships between the two criteria tend to be as high as reliabilities would permit. This may be due to the fact that petty officer qualities and technical competence are actually so intimately associated that differentiation is really impossible. It may well be, for example, that technical competence is the chief determiner of competence as a petty officer; on the other hand it may also be true that an individual will not be given opportunity to demonstrate high levels of technical competence unless he possesses correspondingly high petty officer qualities. If narrow segments of technical competence had been studied (e.g., code receiving or sending in the case of radiomen) it is possible that a lower correlation between petty officer qualities and technical competence would have been obtained.

In future studies it would seem imperative to experiment with the use of performance tests and check lists as criterion measures. This approach was actually recommended for use in the study reported in this chapter; but due to the difficulties of administering such tests and check lists to personnel aboard ships in or preparing for combat operations, the idea was abandoned. It should be pointed out, however, that this method also has serious disadvantages, the chief ones being the differences between ships in types of gear and methods of operation and the differences in jobs performed by men who are in the same rating and pay grade. Only with these factors carefully controlled would the use of tests and check lists be successful. The scores obtained would, of course, have to be corrected for differences in experience in much the same manner as was done in this study.

Relationship Between Background Factors and Shipboard Performance. The relationship between Basic Test Battery scores and the criterion are encouraging. Unfortunately, complete data were available on only about 50 per cent of the cases; hence, it is not possible to make any sweeping conclusions about the relationships. In general, the findings were in line with expectation, agreeing quite well with those found in the case of elementary and advanced

enlisted schools. Of special interest was the finding that the General Classification Test showed a significant correlation with the technical competence of radiomen. Actually this correlation is higher than that found between General Classification Test scores and final grades in radio school. This can perhaps be attributed to the fact that radio schools emphasize ability to learn the code, while aboard ship the emphasis is placed on several factors in addition to code proficiency. Results of this sort if found in large numbers of cases aboard many ship types should be of real significance to classification and training.

The data on the relationship between Navy training and shipboard performance are difficult to interpret. Unfortunately the number of usable cases was small and two ratings had to be omitted from the analysis because of small numbers of cases. While an attempt was made to equate the school trained and non-school trained groups on the basis of ability, there is no assurance that this procedure was completely satisfactory. In short, the findings must be interpreted with extreme caution despite the fact that the findings have some degree of plausibility. It seems reasonable, for example, that school training for radiomen should be more important than for gunner's mates. Opportunities for learning about the job aboard ship appear greater in the latter case than in the former.

One reason for warning against generalization from these data is the fact that they were gathered aboard only relatively large ships —destroyers, cruisers, carriers and a battleship. On ships of this type there is usually a substantial number of experienced men who can fill the more responsible positions and assist in the training of newcomers. In other words, it may not make a great difference whether a man is school trained, when he has to start at the bottom anyway. On a small ship the situation is quite different; new men must fill immediately positions of considerable responsibility. It seems quite likely, therefore, that school training may be very important for personnel being sent to small ships but of only minor importance for those going to large ships with experienced crews. This question should be made the subject of careful study, for there would be important implications in such findings for the assignment of school trained men, especially in wartime. If school training is essential in the case of men going to small ships and of lesser importance for those going to large ships, cognizance should be taken of such a fact in periods such as prevailed during World War II when there always seemed to be too few trained men for the many billets to be filled.

A second reason for recommending caution in generalizing from these data is that the criterion employed falls short of desired stand-

ards. As stated earlier in the chapter, it was the aim of the research to determine the relationship between background factors and the criterion of technical competence. Since a high correlation was found to exist between the technical competence, petty-officer qualities, and overall desirability ratings, there is grave doubt about the acceptability of the criterion. Until other criterion data are available, such as would be obtained from valid performance tests, it would be unwise to formulate any definite conclusions about the effectiveness or non-effectiveness of school training.

The low correlations of age and amount of civilian education with the criterion are not surprising. It should, of course, not be concluded that there is no relationship whatever between these two factors and performance aboard ship. It is important to remember that the age range was primarily from 20 to 30. Within that age range the relationship was very low; but if men up to 50 years of age had been included, it seems likely that the relationship would have been larger. The same principle applies to amount of civilian education. The majority of men included in the study had completed 10, 11, or 12 years of education. If there had been a normal distribution in the amount of education, for example from 6 to 16 years, there probably would have been a somewhat larger relationship. It seems quite likely, however, that if a man's Basic Test Battery scores are available, additional knowledge of how many years of civilian education he completed will not improve the efficiency with which his success aboard ship can be predicted.

It is unfortunate that satisfactory information about occupational experience and personality factors was not available for this study. Differences in the coding procedures of occupational experience or lack of occupational experience on the part of a large share of the younger men, and failure on the part of classification interviewers to record enough information about a man's civilian job, made the study of the occupational factor very difficult. As a result the findings in this study were entirely negative.

While the data for this study were being gathered, the investigators were frequently told that personality factors are very important in determining quality of performance aboard ship. Similar reports were made in a preliminary study which was made of the shipboard performance of junior officers. For many shipboard billets it seems that interest in the work, trustworthiness, and ability to get along with others are the variables which are really associated with success, the degree of technical competence required being no greater than can be acquired by an average individual in a reasonable period of time. If and when these traits can be measured

objectively, it seems likely that the efficiency of prediction of shipboard performance will be greatly improved.

The results of this pilot study indicate that research on the relationship between background factors and performance aboard ship is feasible and that it can provide useful information for classification and training. The two major weaknesses in the study reported in this chapter are (1) the incompleteness of the data in the enlisted personnel records and (2) the inadequacy of the criterion of shipboard performance. The first weakness can be corrected by the systematic testing and classification of all personnel in the Navy; the second will require intensive research over a period of time, leading to the development of more acceptable criterion measures.

CHAPTER XXI

INFORMATION SURVEYS AS EVALUATIVE DEVICES

INFORMATION SURVEYS is the title given to a series of projects designed to determine and analyze the opinions of Navy men about various aspects of training, educational programs, and Navy life. This chapter is concerned with the use of this survey technique for the evaluation of personnel practices. The previous chapter was concerned in part with the evaluation of selection and training procedures by analyzing the relationships between classification and training data and performance aboard ship. In presenting the case for the initiation of opinion surveys, it was pointed out that investigations into the general area of morale and attitudes could provide useful data for the appraisal and improvement of personnel and training policies. What the men think about their training, or the personnel practices which affect their Navy life, may not correspond to the judgment of experts; but it is real to the men themselves. When the experts know the distribution and magnitude of these opinions, they are in a position to take practical steps to modify those which appear to rest upon error or misconception or are for other reasons regarded as undesirable.

The three major opinion surveys which are reported in this chapter represent only a very modest attempt to probe the complex area of attitude and morale. The first survey was not begun until January 1945, and all three surveys were carried out by staff members who also had other duties. Nor do the three surveys constitute a coordinated program of research: each survey had a different purpose, and each, except for the first, was undertaken in response to a specific request. In spite of the different contexts of each survey, there is enough overlapping in content and especially in methodology to permit drawing a number of generalizations both with respect to the attitudes of Navy men and with respect to the values of attitude research in evaluating personnel practices. Before these generalizations are drawn, however, descriptions of each survey, covering methodology, content, the plan of analysis, and major results will be presented.

Survey of Navy Training

Throughout the war the Test and Research Section was located organizationally within the Training Activity of the Bureau of Naval Personnel. Therefore it was natural for the first survey to be directed toward the topic of Navy training. Training had expanded

tremendously; new schools were established, selection procedures were set up, new curricula were written, new instructors trained, new achievement examinations constructed. The experts had worked with all the resources at their command to do the best job they could. Now, what did the enlisted men who were put through these training programs think about them? The first survey, therefore, was to find out what enlisted men thought about their training, particularly the training for their specialty in Navy schools. Called Information Survey 1, the title was intended to convey the notion that the primary purpose of the survey was to give curriculum and instructor training experts information which would be helpful in appraising the strengths and weaknesses of their programs.[1]

How can a survey of opinions provide this information? Of what value is the opinion of trainees? In looking at the educational process, particularly in a military situation, one is tempted to center his attention too exclusively on the instructors, the curriculum makers, and other experts. Training, however, is not a one-way process; it is a two-way process involving both a trainer and a trainee. It is the student who is the learner. Learning is conditioned by the student's opinions and attitudes with regard to the importance and probable usefulness of the material, by his interest in the material, and by the many other elements subsumed in the term motivation. Moreover, the student is the final applier of the learned material. He is, therefore, in a uniquely favored position to look back on his training in the light of the duties he has had to perform.

The opinion of the consumer, as a criterion of training, is of course not equally valid under all conditions. The problem is to get the opinions of those consumers who can look back on their training from the most valid vantage point. Trainees who have not yet completed their training program may have useful and highly suggestive opinions about the way in which the program is being conducted, but they are not in the best position to appraise its value. On the other hand, trainees who have been out on the job for several years may have forgotten a good deal about the training program they went through. Consequently they tend to have only a very general impression of its effectiveness and to have difficulty recalling specific strong and weak points. The man who is faced during his first few months on the job with the necessity of solving problems which his training has purportedly prepared him to solve is in the ideal position to appraise his training. From this position the man's opinions are most likely to have an intensity and a validity of greatest value to the training experts. Information Survey 1, in

[1] The assistance of Lt. Comdr. Duane L. Gibson in developing and analyzing this survey is acknowledged.

other words, may be regarded as a type of follow-up study in which the subject-matter is opinions rather than performance.

METHODOLOGY. In designing Information Survey 1, three basic problems had to be faced. The first was the problem of *content.* What sort of questions about training should be asked? Two types were agreed upon: (1) general questions about training methods applicable to almost any training program, and (2) specific questions about the curriculum content of each specialty to be answered only by men who had been trained for that specialty. The first type included such questions as what the men thought about the usefulness of and the time spent on lectures, laboratory work, discussions, demonstrations, training films, slides, etc. Men answered these by checking such categories as "too much," "about right," "too little;" or "too many," "about right," "not enough;" or "all or most were helpful," "about half were helpful," "few or none were helpful." For the second type of question the combined judgment was obtained of the people who wrote the school curriculum and the people who constructed the final achievement examinations as to what were the major emphases of the training program. These major emphases for each of ten specialty ratings were listed on special one-page supplementary sheets and the men were asked to check "too much," "about right," or "too little" in response to each one in the light of the duties they had had to perform on the job.

The second basic problem was concerned with the *population.* What sort of men should answer the questionnaire? Ten special training programs were selected—the ones for which special one-page questionnaires were developed. These ten specialties (electrician's mate, fire controlman, gunner's mate, machinist's mate, motor machinist's mate, quartermaster, radarman, radioman, signalman, torpedoman's mate) accounted for over sixty per cent of all rated men aboard ship. It was considered important also to obtain opinions of recent trainees, men who had received their school training not more than a year prior to the study. Only those men who had actually worked in their specialty would be analyzed.

The third basic problem, and by far the most difficult, was the problem of *field work*—where to find the sort of men wanted. From the central office in Washington two types of location looked suitable —shakedown centers and redistribution centers. The latter proved to be relatively unprofitable because most of the men returning from sea duty were not recent trainees. The shakedown centers, and also Navy yards, were good for testing men aboard destroyers, cruisers, battleships, and carriers. For smaller ships no very satisfactory way was found to handle the field work.

The sampling plan was to give the questionnaire to groups of

men in those places where a reasonable proportion of cases which would fit the sampling objectives could be expected. After the data were punched on IBM cards, selected cases would be sorted out for analysis. For example, in appraising training in special schools, only those men would be used who (1) had attended a naval training school for one of the ten specialties immediately following recruit training, (2) had graduated from that school, and (3) had worked on a job which was in their specialty. The method, in other words, was to set general objectives for the survey, and then, because there was no effective control over the field situation, to select the sample of cases for analysis after the survey was administered. Using this approach, about sixty per cent of the questionnaires proved to be usable.

RESULTS. The first group of results to be reported centers in the area of training methods in Navy schools. Only men who had graduated from a Navy school and had worked in the field of their specialty are included in this analysis.

Training Methods. Replies to an overall appraisal question were distributed as follows:

> *Question:* In the light of what you now know about the duties of your specialty (rating) what do you think about the general value and usefulness of your school training?

Reply	Percentage of graduates
All or most of it was important and useful	64
More than half of it was important and useful	18
Half or less of it was important and useful	18

Answers to specific questions about training methods were then studied for each of the three groups noted above. Among many significant relationships thus revealed, two stood out more clearly than the rest.

Among men whose general appraisal of training was	Percentage who said that the equipment studied in school was very different from the equipment used on the job
high	15
medium	20
low	46

Among men whose general appraisal of training was	Percentage who said that all or most instructors were able to put across their material successfully
high	61
medium	53
low	32

The problem of coordinating equipment between the schools and the fleet is an administrative one. The problem of providing good instructors for the schools is primarily a selection and training problem.

The distribution of replies showing the opinions of men with regard to their instructors is indicated below.

Question: How many of your instructors were able to put across successfully what they were trying to teach?

Reply	Percentage of graduates
All or most of them	54
About half of them	36
Very few of them	10

In order to identify what specific practices or elements in the training program were associated with high appraisal of instructors, the total sample was divided into the three groups indicated above, and the responses of each of these three groups were analyzed. This analysis revealed that the "good teacher," more than the poor one, allowed more time for practice in operating equipment, did not spend too much time on lectures, and used demonstrations and discussions more frequently. The good teacher, also, used films, slides, demonstrations, and discussions more effectively than the poor teacher.

The main implications from these analyses are that training can be improved, in the opinion of these men, by providing:

1. Better instructors
2. Equipment more nearly like that in the Fleet
3. More time to operate equipment
4. More time in the shop and laboratory
5. More demonstrations
6. More time for questions and discussion

When these questions about training methods were studied for each of the ten specialties, several differences in responses were noted. For example, among men from signalman, radioman, fire controlman, and motor machinist's mates training programs, 60 per cent or more said that all or most of their instructors were able to put across successfully what they were trying to teach. But in the machinist's mates, quartermaster, and radarman programs, less than 40 per cent expressed this opinion of their instructors. Similarly, when the overall appraisal of school training was tabulated for each specialty, it was noted that less than 10 per cent of the men from gunner's mates, radarman, and radioman programs gave a low ap-

praisal, whereas more than 25 per cent of the men from fire controlman, machinist's mates, quartermaster, signalman, and torpedoman's programs gave a low appraisal to their training.

SPECIAL CURRICULA. Two examples of responses to the special one-page questionnaires developed for each of the ten ratings are reported next. For each rating the range of responses was wide, suggesting that the men responded differentially to the various items. To illustrate, in the questionnaire for signalmen as few as 12 per cent of the men said they had "too little" training in sending and receiving semaphore messages; but 41 per cent expressed the opinion that they had "too little" training in flaghoist spotting, bending on and stowing, and 70 per cent felt they had "too little" training in sending messages by flashing light. In the questionnaire for electrician's mates, as few as 18 per cent of the men reported "too little" training in the use of hand tools and in electrical theory and mathematics in contrast to 59 per cent who felt they had "too little" training in circuit tracing and in blueprint reading and mechanical drawing. Results such as these are suggestive of the strong and weak points in the training curricula, as judged by the men who have had the training and have applied it to their jobs in the fleet.

Survey of Educational Services

The second opinion survey, called Information Survey 2, was designed to appraise the Navy's educational services program. The educational services program (1) provides opportunity for both officers and enlisted men to attend voluntary off-duty classes, take correspondence courses or obtain textbooks from the Armed Forces Institute, (2) distributes such materials as newsmaps and war orientation readings, (3) promotes talks and discussions about the background and progress of the war, and (4) carries on similar enterprises.

Prior to the war, the Navy's orientation program was relatively modest. With the influx of inductees, however, and the growing interest in understanding national and world events, the Navy expanded its orientation program and added many new features. Moreover, the fact that many young men were taken directly from schools into the Armed Services resulted in the Navy's having a type of population with which it had not previously had to deal in such large numbers. In cooperation with the colleges and public schools, both the Army and Navy developed methods whereby service personnel could maintain contact with and continue the pursuit of their civilian educational interests. As is true of all new enterprises, this educational services program was regarded with skepticism by

many officials both in and out of uniform. It was felt, therefore, that a survey revealing the extent of participation in the program, the characteristics of the participants, and the values claimed by the participants would be helpful and timely. Accordingly, the Educational Services Section requested that a survey be made.

METHODOLOGY. The educational services program is Navy-wide, and for this survey random samples of the total Navy population were wanted. For these random samples the Educational Services Section wanted particularly a good representation of men from overseas bases and aboard ships. Questionnaires and instructions for their administration were sent to eight overseas bases; and the men were selected by one of two methods, either by calling in all men whose service numbers ended in a given combination of figures, or by selecting every *n*th name from a roster of the ship's company. The bases included large ones and small ones, old ones and new ones, and were spread geographically from the Aleutians to the South Pacific. The shipboard sample was obtained by administering the questionnaire to large unselected groups of men returning from sea duty. These men were tested during their first day at receiving stations (most on the West coast and a few on the East coast) and were instructed to answer the questions in terms of conditions aboard the ship on which they had just served. An additional sample was obtained from the ship's company of one of the large training centers in the United States. Ninety per cent of all the questionnaires sent out were returned.

The design of the questionnaire reflected the major aspects of the educational services program and included three types of questions: (1) questions on participation in the various educational services activities, (2) questions on opinions of or value of these various activities, and (3) groups of questions to measure respectively, information about the war, attitudes about the war and the future, and post-war educational plans. The information questions were scored on the basis of the number of questions answered correctly. The score on the attitude questions was the number answered positively. For the estimate of post-war educational plans two approaches were used. One was simply the number who said "go to school" in answer to a free-response question about the sort of work they planned to do after the war. The other was the number who, in answering three different questions about post-war plans, consistently checked the category indicating an intention to go to school. By scoring these questions on information, attitudes, and post-war educational plans, it was possible to relate each of these factors, when appropriate, to participation in educational services activities and opinions about educational services enterprises.

RESULTS. The results were presented in a series of nine reports under the general heading *Navy Men Appraise Educational Services.* These results are summarized here under the two general headings of orientation and education.

Orientation. The first series of results is described under the general heading of orientation. Most Navy men said that they got the news every day; 61 per cent reported daily reading of a paper and 71 per cent reported daily listening to radio news broadcasts. Eighty per cent of the men said that the Navy's weekly newsmaps were available at their location. Interest was also high in having talks and discussions on the background and progress of the war; 60 per cent indicated that they would like to hear regular talks at least once a week. The questionnaire also included a list of twelve topics related to the war, its background, and the nations involved, and asked the men to check each topic they would like to read about during off-duty time. Every topic was checked by more than half the men, and two-thirds of the men expressed interest in reading about seven or more of the twelve pamphlet topics listed.

The relationships between these expressions of interest (in news, talks, and readings) and certain expressions of attitude are shown below.

Among Men for Whom a Clear Understanding of Why We Are Fighting Is	Percentage of Men Who Want Talks and Discussions at Least Once a Week	Percentage of Men Who Get News Daily		Percentage of Men Who Are Interested in 11 or 12 Orientation Readings
		By Paper	By Radio	
a. Absolutely necessary	45	64	73	42
b. Of great importance	41	61	72	34
c. Of medium, little or no importance	25	54	62	19

Among Men Whose Attitudes about the War and the Future Reflect	Percentage of Men Who Want Talks and Discussions at Least Once a Week	Percentage of Men Who Get News Daily		Percentage of Men Who Are Interested in 11 or 12 Orientation Readings
		By Paper	By Radio	
a. High motivation	46	67	76	44
b. Medium motivation	40	59	71	33
c. Low motivation	34	56	64	25

The figures indicate that men who say it is absolutely necessary to them personally to have a clear understanding of why we are fighting show the highest level of interest in talks and discussion, news getting, and orientation reading.

The same pattern of relationship is indicated between these expressions of interest and attitudes about the war and the future.

Four questions in the survey were grouped to obtain a measure of these attitudes.

"In your opinion is the United States fighting for things that *you* feel are worth fighting for?" (check one)

1__ Yes
2__ I think so, but I'm not sure
3__ No
4__ Undecided

"Do you ever get the feeling that this war is not worth fighting?" (check one)

1__ Very often
2__ Sometimes
3__ Only once in a great while
4__ Never

"After the war is over, do you think the United States will be a better place or a worse place to live in than it was before the war?" (check one)

1__ *Worse* after the war than before the war
2__ *Same* after the war as before the war
3__ *Better* after the war than before the war
4__ Undecided

"How important do you think it will be for the United States, Russia, England, and China to cooperate closely with each other *after* the war?" (check one)

1__ Absolutely necessary
2__ Pretty important
3__ Not so important
4__ Don't know

These four questions were "scored" and the men grouped according to the number of questions which they answered "positively." Those who doubted that the U. S. was fighting for worthwhile goals, who sometimes or often felt that the war was not worth fighting, who thought the U. S. would be worse after the war, and who doubted the necessity of close cooperation among the major allies after the war were regarded as having "negative" attitudes. Men who held positive convictions on all four questions were regarded as having "high motivation"; men who held positive convictions on three of the four questions were regarded as having "medium motivation"; and men whose attitudes were positive on only two, one, or none of the questions were regarded as having "low motivation."

Interest in orientation—news, talks, discussions, and readings—is high. Opportunity to satisfy this interest should be equally high. The survey indicated several major areas where opportunities should

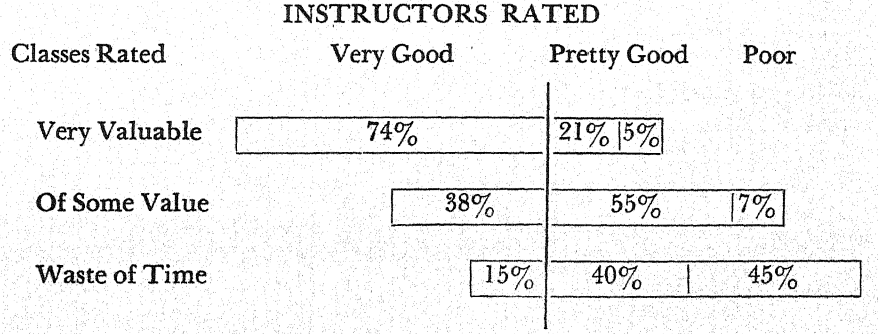
be widened. Among men afloat, for example, the proportion who got the news daily and who saw the newsmaps weekly was significantly smaller than the proportion among men ashore. Also, the interests of only a few men in talks and discussions were being satisfied: among men who would like to hear talks at least once a week, only 28 per cent actually heard talks that frequently, and 54 per cent of these men said they never heard any talks.

With respect to orientation, in summary, the survey demonstrated that the interests of men were high, that these interests were related to positive convictions about the war and the future, but that opportunities to satisfy these interests were limited.

Education. The second major area of the survey is considered under the general heading of educational opportunities. The results indicated that 30 per cent of all naval personnel were using educational opportunities offered by Educational Services—either attending voluntary off-duty classes or taking work under the United States Armed Forces Institute (USAFI) program.

One man in five said he had attended off-duty classes at his present location. Those who participated belonged to no special group as to age, length of service, or rank. They were a fair cross section of the Navy as a whole. A significantly higher per cent of high school graduates did, however, attend the classes—60 per cent of those who attended were high school graduates in contrast to 51 per cent graduates in the total sample studied. Opportunities offered in the way of voluntary classes and attendance varied from place to place: at advance bases 72 per cent of the men said off-duty classes were offered, and 19 per cent of the men attended. At the training center 82 per cent said there were off-duty classes; 21 per cent attended. Among men afloat, chances for attending classes were most limited; only 22 per cent said there were classes, but 10 per cent attended them.

In rating the off-duty classes, 42 per cent of the men said they were very valuable. Only 5 per cent considered them a waste of time. And, significantly, the way men rated their classes corresponded with the way they rated the teachers of those classes.



One man in five participated in the work of the United States Armed Forces Institute which offers correspondence courses, self-teaching courses, and textooks for self-study. The survey showed that 10 per cent of the men had signed up for a correspondence or self-teaching course from USAFI, 5 per cent had gotten textbooks, and 3 per cent had done both. Of the men who took work with USAFI, 47 per cent rated their study as very valuable, only 2 per cent as a waste of time.

Post-war jobs and personal interest were the chief motives for study. Those who participated in the USAFI program or who attended off-duty classes gave these reasons:

Reasons	Percentage of Men Giving Reasons for Participation in USAFI Work	Off-Duty Classes
Subject was related to work I hope to do when the war is over	28	16
Wanted to learn something which interested me personally	18	26
Subject was related to my Navy Job	12	25
To get high-school or college credit	25	10
All others	17	25

The desire to learn more about their Navy jobs and the hope of getting high-school or college credits were almost equally important motives for study.

Of the 5 or 6 per cent of Navy men in the survey who indicated definite plans to return to full-time school after the war, one in five had already applied for high school or college credit for his Navy training. Even men who did not say that they intended to return to school were interested in the procedure the Navy has worked out for accreditation. Of all the men surveyed, 25 per cent had asked advice about school credits and courses, and 7 per cent had actually applied for credit. Among men who had applied for credit and had received an answer from the school, 78 per cent received a favorable answer.

The wisdom of providing opportunities for voluntary study is amply confirmed by the substantial proportion (30 per cent) of Navy men who were found taking advantage of those opportunities. While men of all sorts were making use of these services, they had a special appeal for men whose formal education had been interrupted by the war. The orientation services were found to be operating effectively with respect to keeping men up-to-date on the news of the war. Opportunities to discuss and clarify some of the important background and causal events and ideas related to the war were,

however, more limited. Most men felt that it was highly important to them personally to have a clear idea of why they were fighting. Partly as a result of the survey, the Navy expanded its Educational Services program, laying greater emphasis on the formation of discussion groups and increased opportunities for study during the post-hostilities period.

Attitude of Amphibious Forces Personnel

The third major survey was concerned with the attitudes and opinions about the amphibious forces which were held by enlisted men in that type of duty. The purpose of the study was to determine the opinions which men have about amphibious duty and to identify factors which were related to favorable opinion so that positive steps might be taken for improvement of attitudes and opinions. Amphibious warfare was largely a development of World War II. Because of its urgent need for large numbers of personnel and because of the necessarily rapid expansion which it underwent, the amphibious forces, were assigned an especially heterogeneous assortment of personnel. Requiring a mass base, amphibious duty, unlike naval aviation or the submarine service, was not generally regarded as an elite or prestige-bearing type of duty; yet its importance in winning the war was indisputable. Also, because of its newness and its size, training methods were less well established than for older and more exclusive types of duty. This combination of circumstances made it desirable to conduct an analysis of attitudes and morale among amphibious forces personnel.[2]

METHODOLOGY. The survey was initiated at the request of the Amphibious Training Command, Pacific Fleet. A questionnaire was administered during the months of April and May 1945 to a class of about 800 men who had just completed primary training for small attack boats and to 400 men who had just completed advanced training for these small boats. In addition, in order to provide a sample of men with amphibious experience for comparison with those recently trained, the questionnaire was administered to 400 men in ship's company and attack boat crews of four attack transports.[3]

The questionnaire was developed after extensive analysis of similar studies conducted by the Research Branch, Information and Education Division, of the Army General Staff and in consultation with personnel at the Amphibious Training Command.

[2] The development and analysis of this study was guided by Mr. George Hausknecht with the assistance of Lt. Comdr. Duane L. Gibson.

[3] Personnel comprising the attack boat crews on board attack transports (APA's) are not considered to be members of the regular ship's company manning the attack transport.

In order to avoid giving the impression to the men questioned that the amphibious forces were being singled out for special study, the questionnaire was drawn up in such a manner as to make it applicable to persons in any type of Navy duty.

In addition to the usual questions about the background and extent of naval experience of the men in the sample, the questionnaire included items to obtain information about each of the following areas of opinion: (1) fundamental motivation, (2) personal preference for type of duty and opinion of the prestige of different types of duty, (3) satisfaction with job assignment, (4) attitude toward officers, (5) opinion of shore training, (6) personal adjustment, (7) opinion about specific aspects of amphibious duty.

The questions in each of these areas were subjected to an empirical analysis to determine which combination of responses would give a meaningful and practical description of each area. It was possible to obtain certain indications of the magnitude of the problem through the study of the distribution of responses. The relation of opinions in each area to attributes of the men and to various aspects of Navy life was analyzed in order to suggest personnel policies which might be helpful.

RESULTS. The findings of the study were presented in five mimeographed reports, each covering a related group of attitudes.

Fundamental Motivation. By combining responses to the three questions noted below, it was possible to arrange the men in the sample into motivation types.

"In your opinion is the United States fighting for things you feel are worth fighting for?"

1___ Yes
2___ I think so, but I'm not sure
3___ No
4___ Undecided

"If it were up to you, what kind of an outfit would you rather be in?"

1___ In a combat outfit overseas
2___ In a non-combat outfit overseas
3___ In an outfit that will stay in the United States

"If it were up to you to choose, do you think you could do more for your country serving the Navy or as a worker in a war job?"

1___ Serving in the Navy
2___ As a war worker
3___ Undecided

The combined tabulation of these three questions resulted in a grouping of the men in six motivation types in terms of (1) ac-

ceptance of hardships and risks in support of their principles, (2) acceptance of the principles but avoidance of the consequences, and (3) disbelief in the principles. The six motivation types and their distribution among the sample of amphibious forces men were as follows:

Group	Responses to Three Questions	Percentage of Men in Each Group
A	Feel the war is worth fighting; choose combat; choose military service	22
B[4]	Feel the war is worth fighting; would avoid service; but do choose combat	14
C	Feel the war is worth fighting; choose service; but would avoid combat	16
D	Feel the war is worth fighting; but would avoid military service and combat	35
E	Doubt the worth of the war; would avoid military service and combat	10
F	Do not think the principles are worth fighting for; would avoid military service and combat	3

Despite the rather large proportion who would prefer not to be in combat duty (about two-thirds), the response of the men to questions on their confidence in their ability to perform adequately in combat was highly favorable. Fewer than 10 per cent of the men felt that they would never be able to do well in combat.

Even among the best motivated men, the proportion of those who would avoid amphibious duty if they could was seriously high, with nearly half saying that they would try hardest to stay out of the amphibious forces over all other types of Navy duty. Among the most poorly motivated group, 69 per cent would avoid amphibious duty.

Motivation showed a marked relation to personal adjustment to life in the Navy and to convictions about the outcome of the war, both as to its effect upon life in the United States after the war and the likelihood of achieving the goals for which the war was fought.

Prestige of Amphibious Duty. Amphibious duty was found to be held in higher esteem by men aboard amphibious vessels than by men at training centers. A significantly larger proportion would freely choose it as a preferred duty and far fewer would avoid it among the men aboard ship as compared with the men at training centers. In expressing their preference for a type of duty, men in the shipboard sample placed amphibious duty first, aviation second, and destroyers third; but among men at training centers, aviation

[4] This group, which appears inconsistent, is interpreted as men who would prefer to avoid military service, but since they are now in the Navy they accept the idea of combat duty.

was the most frequent choice, destroyers second, armed guard third, with amphibious duty in ninth place.

The low prestige which is accorded to amphibious duty by men at the training centers may well be partially accounted for by lack of information or negative information which is, to some extent, dispelled after assignment to actual duty aboard ship. A majority of the men claimed to have heard nothing about the amphibious forces before assignment; but among those who had heard anything, the great majority had heard unfavorable rumors.

In reply to a question as to whether respondents had changed their attitudes toward amphibious duty, about two-thirds of the men afloat and at training centers indicated that they had changed their views. Ninety-two per cent of men afloat changed in the direction of greater respect for such duty, as compared with 61 per cent for those in training centers.

Men were asked to indicate a preferred duty and to explain their preference. They were also asked which duty they most wanted to avoid and to explain their reasons. These reasons were in support of choosing or avoiding various duties, not necessarily amphibious duty. They form a generalized set of factors in the prestige of any duty.

Only two positive factors (i.e. elements of prestige causing men to choose a particular duty in preference to all other duties) were mentioned by as many as 10 per cent of the sample. These were "importance in winning the war" by 33 per cent, and "chance to use skills and training" by 10 per cent. Four negative factors in prestige were offered by 10 per cent or more of the sample. These were: "high casualty rate," 18 per cent; "no advancement," 16 per cent; "excessive hardships," 13 per cent; "low service prestige," 12 per cent.

An analysis of the place of these positive and negative factors in prestige as they relate to amphibious duty showed that more than half of the men felt that this duty was most important and about half felt that little or none of their work could be done by men with less training and ability, indicating a feeling of pride in the skills which they used. But "importance in winning the war" appears in many ways to be an indirect way of saying that in their opinion a duty has high public esteem, and half of the men in the sample felt that the public did not know enough about the amphibious forces. With respect to the negative prestige factors, 4 in 10 of the men in the sample felt that amphibious duty was the most dangerous; 6 in 10 that it involved more hardships than any other duty, and 5 in 10 that the rest of the Navy looked down on amphibious forces personnel.

The responses to many of these items pointed to the need for an active indoctrination program which would provide correct information about the factors which men consider important in the prestige of their duty. About half of the men expressed an interest in additional orientation in the amphibious forces, including one-fourth of the total stating that they had never had any.

Attitude Toward Officers. By combining responses to two questions, the men in the sample were separated into officer appraisal types. These were based on the men's opinion of the proportion of officers who are personally interested in them and their opinion of the willingness of those in charge to help. About half of the men in the sample felt that they could get help most of the time from at least a few officers who took a personal interest in them.

Men's experience in the Navy appeared to influence greatly their opinion of officers. Men in a training situation were found to have a much lower opinion of officers than men aboard ship. Unfavorable opinion of officers was also found more frequently among men who had been in service for some time than among those with shorter naval experience. Men sent into training for amphibious duty after a year or more of Navy service were extremely unfavorable in their attitude toward officers.

Officer Appraisal Type	Response	Percentage of Men Giving Each Response at Training Center	Ship
5	All or most take a personal interest and help all or most of the time	22	36
4	Half take a personal interest and help all or most of the time	9	14
3	Few take a personal interest and help all or most of the time	13	19
2	Half take a personal interest but they often don't help	5	14
1	Few take a personal interest and they often don't help	25	15
0	Few or none take a personal interest and they almost never help	26	14

Attitude toward officers was not related to such personal characteristics of the respondents as age, education, marital status, pay grade or specialty. On the other hand, several aspects of naval experience in addition to those of training and length of service were related to opinion of officers. Men who felt that their officers had a personal interest in them and were willing to help them were more likely to have a favorable opinion of amphibious forces duty and to be satisfied with the jobs which they were doing within that duty.

These men with a favorable opinion of their officers were also more likely to be well motivated about the war and less likely to have symptoms of maladjustment to Navy service.

It should be pointed out that there are doubtless many instances in which the claim that officers are not personally interested in their men is unjustifiable. If, however, the men *feel* that there is a lack of interest, it is *real* to them and thereby becomes a problem which must be dealt with.

Job Satisfaction. Job satisfaction was measured by a score obtained from the combined responses to two questions:

"How satisfied are you about being in your present Navy job instead of some other?" (check one)

1___ Very satisfied
2___ Satisfied
3___ It doesn't make any difference to me
4___ Dissatisfied
5___ Very dissatisfied

"Would you change to some other Navy job within your type of duty if you were given the chance?" (check one)

1___ Yes
2___ No
3___ Undecided

The combined responses to these questions gave fifteen categories of job satisfaction. These were reduced to five for purposes of simplicity in analysis. The largest area of dissatisfied men were in training centers, where jobs are less well defined and hardships often appear less reasonable than at sea. The proportion of men in training and afloat in each job satisfaction category is shown below.

Score	Job Satisfaction Category	Percentage of Men in Each Category at Training Centers	Ship
4	Satisfied and would not change	6	30
3	Satisfied but undecided about changing	3	8
2	Contradictory responses or completely undecided	14	25
1	Dissatisfied but undecided about changing	3	3
0	Dissatisfied and would change	74	34

Personal attributes such as age, education and marital status were not closely related to job satisfaction. There was some relationship between job satisfaction and pay grade, but only for men on board ship.

Job dissatisfaction was also closely associated with poor personal

adjustment to Navy life and with a negative attitude toward amphibious duty. The index of adjustment to Navy life was obtained from the pattern of replies to three questions included in the survey. Only 3 per cent of the men who were satisfied showed a low personal adjustment score, compared with 22 per cent of the dissatisfied. Among men who would choose amphibious duty, 7 per cent were dissatisfied with their jobs, as compared with 81 per cent dissatisfied among those who would like to have rejected it.

Certain attitudes relating to the job also appear to be important in contributing to job satisfaction. The following table identifies these attitudes.

Attitudes	Percentage of Men Who Were Satisfied	Percentage of Men Who Were Dissatisfied
Were very much interested in their jobs	75	13
Believed they had a good chance to show their abilities	72	15
Believed everything possible had been done to place them in the best possible job	56	4
Had civilian skills which they believed were being used at least half the time	48	13
Thought they had a good chance for promotion	37	12

Apparently job satisfaction can be improved by taking positive steps to improve these elements of job satisfaction. Greater effort directed toward a program of orientation both to Navy life and to amphibious duty would also contribute to a better attitude toward the job.

Opinion of Shore Training. The enlisted men's evaluation of their shore training was based on the following questions, asked separately for those with sea duty and for those who had not yet been to sea:

(For those who had had sea duty)
"Do you feel that your training ashore made it possible for you to handle most things that came up?"
(For those who had not had sea duty)
"Do you think your training will make it possible for you to handle most things that come up?"

Reply	Percentage of Men
1. "Yes, my shore training gave (will give) me the 'know how' for almost everything"	14
2. "Yes, but I had (will have) to pick up a few things on my own"	46
3. "I was (will be) about half trained"	8
4. "No, I had (will have) to learn a great deal on my own"	24
5. "No, the training didn't (doesn't) amount to anything"	8

For purposes of analysis, men who checked the first or second replies were considered as having expressed the opinion that their training was "adequate." Those who checked one of the last three responses (i.e. "half trained" or less) were considered as indicating the feeling that their training was "inadequate."

The more recent the training, the higher the opinion of its adequacy. About two-thirds of the men who had been in the Navy less than one year felt that their training was adequate, compared with little more than one-fourth for those who had been in for more than three years. Likewise, men who had up to six months of sea duty considered their training more adequate than men who had over two years of sea duty. The most likely explanation of this lies in the fact that men who had already served over two years at sea had received their shore training during a very early stage of the war when, admittedly, systematic training for amphibious duty was just being developed.

Appraisal of training was found to be so interrelated with other aspects of adjustment to Navy life as to appear to require a simultaneous attack on several fronts in order to effect major improvements. Men who felt that their shore training had been adequate were significantly more likely to be well motivated, well adjusted to Navy life, to feel favorably toward their officers, to have increased respect for the amphibious forces, and to be high in other indications of good adjustment to Navy life.

Half the men who considered themselves adequately trained stated that their instructors had plenty of experience and could get the teaching across. This compares with 27 per cent for those who felt themselves inadequately trained. Among those who felt they were adequately trained, more than half felt that all of their training was necessary, as compared with 37 per cent for those who felt that they had not been well trained. Men who considered their training adequate were more likely to have been provided with orientation in the history and importance of the amphibious forces than were those who felt that they had been inadequately trained.

Summarizing the amphibious forces study as a whole, it may be said that several attitude areas were identified and studied with respect to their association with each other and with numerous specific factors of experience, background, and status. These analyses suggested, in several instances, specific personnel practices or policies which should be instituted or reviewed. The basic fact that most of the attitudes studied were interrelated suggests that the concepts of morale, personality, and adjustment have many specific elements. The improvement of any one specific element, such as providing better indoctrination in amphibious duty or more effec-

tive instructors in amphibious training, may be expected to make only a limited contribution to the overall problem. The analyses suggest that specific practices should be looked upon as related to large problems.

Some Factors Influencing the Value and Usefulness of Opinion Research

The value of opinion research, as illustrated by the three surveys just described, depends on many factors—some technical, some administrative, some judgmental. It would be possible to discuss in detail a series of technical problems, such as sampling, question wording, cross-tabulations, statistical significance of sampling deviations or percentage differences, etc., which relates to the reliability and validity of opinion data. Then, such administrative problems as the interest of management in the results, the appropriateness of the research method to the topic, the use and interpretation of results, could be considered. Rather than discuss a series of specifics, however, we shall consider the more important problems under two broad headings—the problem of suitability, and the problem of interpretation. These categories are arbitrary, but they will serve to stress what the writer regards as the two most fundamental considerations in opinion research, particularly in the Navy.

THE PROBLEM OF SUITABILITY. The suitability of a topic for administrative action is a problem which should be considered in deciding whether or not to conduct a survey. In general, some administrative action should be possible and contemplated, depending on the results of the survey. For example, a survey of what men think of their Navy school training in the light of their subsequent job experience, revealing what the enlisted men regard as strengths and weaknesses of that training, can identify areas of training and specific aspects of training which the experts should re-examine. This re-examination can then lead to changes in the training program. The stimulus for Information Survey 2 was a contemplated administrative action by the Educational Services Section. Authorization to expand its services was about to be granted in anticipation of the need for an enlarged program in the post-hostilities period. An appraisal of the present program was regarded as offering guide-lines for the proposed expansion. In the amphibious forces study, the existence of a morale problem was recognized by certain officers at the Amphibious Training Command. The survey was made to identify this problem more specifically and to analyze it in such a way that improvements in personnel and training practices could be suggested.

The extent to which the topic of a survey is suitable for administrative action is influenced by two questions: (1) does some administrative agency desire the survey, and (2) does the survey permit an analysis of attitudes in relation to personal characteristics and attributes so that a basis for change in personnel or training practice can be suggested? The maximum value to administration from a survey will be obtained when both these questions can be answered positively.

A second consideration is the suitability of a topic or problem to the research methods of opinion surveys. There are a good many questions to which opinion or information surveys can not supply the most useful answers. For example, to ask a cross section of men about their food preferences is not as objective a method of analyzing such preferences as the method of measuring quantitatively the amount of left-overs. Other topics may be considered as properly reserved for expert judgment: for example, should the Navy have more or fewer aircraft carriers? should an advance base be built at X location? should mastery of a foreign language be required of Naval Academy graduates?

A third consideration closely related to the second is whether or not the topic is within the experience of the men to be polled. Do the men have some basis for an opinion? The topic itself may be unsuitable or the kind of questions asked may be unsuitable. Whether Naval Academy men should or should not be required to study a foreign language is a decision for experts. But it would be appropriate to poll Academy graduates to find out how frequently and for what purposes they have had occasion to speak or read the foreign language they studied. In Information Survey 1, the focus of the questions about training was toward the subsequent experience of the graduates: as a result of your experience to date what do you think about various aspects of the training? In the amphibious forces study men were asked about their jobs, their officers, and similar matters with which they had had direct association. Sometimes, of course, it is desirable to ask rather general questions for the purpose of revealing prejudices, misconceptions, or stereotyped opinions. But as a guiding principle, questions which are related to the personal experience of men in the sample will elicit the most useful catalogue of opinions.

THE PROBLEM OF INTERPRETATION. Single questions are usually not very satisfactory indications of complex attitudes. Interpretation of results is often aided, therefore, when several questions are each related to a single general attitude. Such a group of questions may be called a scale. It can be determined experimentally whether a group of questions actually hangs together; or it can be decided by

experts which questions should be studied in relation to one another. If there is such a pattern, the responses to the combination of items can be scored or "scaled." In the amphibious forces study, the responses to two or more questions were combined to obtain opinion scores in the following areas: motivation, personal adjustment, appraisal of officers, and job satisfaction. In Information Survey 2, items were combined to obtain scores on the subjects of attitude toward the war and future, knowledge of war events, knowledge of home-front events, knowledge of background facts, post-war education plans, interest in talks and discussions, and interest in war orientation readings.

When a single general question is asked to obtain a summary opinion about some rather broad topic, the interpretation of responses to it is meaningful when it can be studied in relation to the specific elements it subsumes. For example, Information Survey 1 included a question to elicit an overall appraisal of school training. The replies were then studied in relation to specific aspects of training to determine which were most highly related to the overall appraisal. Similarly, appraisal of instructors was studied in relation to a series of specific teaching practices to determine which practices were most highly associated with high and low regard for teachers.

Another type of relationship which should be considered in interpreting the results of opinion surveys is the relationship among several areas of opinion. The Amphibious Forces Study showed clearly that opinion of officers, job satisfaction, opinion of the prestige of amphibious duty, motivation for war service and combat, opinion of amphibious training, and personal adjustment to Navy life were all related to one another. The implication of such findings is that administrative changes in policy or practice in any one of the areas should be made after estimating the probable effect on related areas. The desirability of a coordinated attack is implied.

A final and highly important problem in interpreting attitude studies is the problem of what the data imply for administrative action. One contribution of an attitude study is to reveal problems or topics which are in need of critical reappraisal by experts. Another contribution is to estimate the effects of personnel and training policies. Do attitudes change following changes in personnel or training practices? The most profitable way to clarify the implication of opinions for administrative action is to analyze them in relation to the attributes of men holding them and in relation to other associated factors. With respect to orientation for amphibious duty, it was shown that certain specific misconceptions were held by many men, that most men had not heard much about the amphibious forces before they were assigned to it, that among those who had

heard something, most had heard unfavorable rumors, and that a substantial number of men expressed an interest in hearing talks, reading articles, or seeing movies about the development, role, and importance of the amphibious forces. The need for a more extensive and effective indoctrination program along definite lines was clearly indicated. On the other hand, the fact that men said, as they did in Information Survey 1, that "too little" time was spent on practice in operating equipment during their Navy school training does not automatically mean that more time should be allotted to this activity. It does suggest, however, that experts should reappraise the distribution of time in training schools in the light of the opinions expressed by men who have been faced with the necessity of applying in the Fleet the knowledge and skill which their training purportedly gave them.

When adequate attention is given to these basic problems of suitability and interpretation, in addition, of course, to the technical research problems, opinion studies can provide valuable and useful evaluative data for personnel and training programs.

CHAPTER XXII

PROBLEMS FOR FURTHER STUDY

In the preceding chapters there has been reported some of the research on Navy personnel problems conducted by the Test and Research Section, NDRC Project N-106, and the College Entrance Examination Board. Aptitude and achievement tests prepared or officially approved by the Test and Research Section have been described, and their use and effectiveness indicated. A perusal of these chapters reveals that although psychological and educational tests were developed and research studies were carried on for a large number of officer and enlisted programs, only a beginning had been made in studying problems and developing procedures for use in certain areas. In other aspects of Navy personnel programs, the needed studies, though contemplated, had not even been started.

The exigencies of the war necessitated work first of all on the most pressing problems—those involving large numbers of persons or particularly difficult or technical training programs or duty assignments, and those which promised the most immediately useful results. As a consequence, the time spent in validation of tests was all too short, the research studies in the field of selection and classification were fragmentary, and studies on training were barely initiated.

By August 1945, however, data were accumulating on the basis of which some evaluation could be made of existing classification and training instruments and procedures, and proposals could be made for trying out other techniques, for developing other possibly more effective instruments, and for examining some scarcely touched but extremely critical problems. This chapter will outline some of the modifications in classification and training procedures which have been indicated by the experience of the Test and Research Section, it will suggest some of the lines of investigation which may be found fruitful, and it will comment briefly upon problems of a methodological nature. No attempt will be made to present an exhaustive or comprehensive program for the post-war personnel research program, since that is the province of the technicians responsible for administering and implementing such a program. There are included only some selected proposals which have grown out of the experience of the past four years and which the writers believe would be immediately useful to the Navy.

Basic of course to any program of selection, classification, and training is sufficient accurate information about what men do in Navy jobs. The job analyses necessary to obtain this information will

need to consider in detail the specific duties, responsibilities, and capabilities involved in each type of job, the degree of proficiency and physical fitness required to perform the various activities, and the extent to which there is overlapping among billets in terms of duties performed or skill and knowledge demanded. Following such job analyses of specific billets, the billets need to be classified, using as small a number of job families or types as is consistent with actual differences in the operation or personnel requirements of the billets. For effective classification and training, these requirements should be expressed in terms of such factors as aptitudes, skills, interests, education, experience, and physical and personal characteristics. Techniques for weighting and evaluating the factors for different types of billets must be developed and validated.

Research on Techniques of Selection and Classification

The goal of a research program in selection and classification should be to provide information and techniques which will make possible the most effective use of the manpower available. Such effective use of manpower implies not only that every officer and enlisted man is placed in the job where he can make his greatest contribution, but also that the optimum placement of each man is accomplished in the minimum amount of time. Given sufficient time, even with inadequate methods of selection, men would probably "shake down" or gravitate to billets for which they are qualified. In the early stages of an emergency, however, when the size of the Navy is increasing rapidly, the saving of time in making suitable assignments may increase considerably the manpower available for combat. One aim of a peacetime research program should be to evolve and put into actual use procedures which can readily be employed in the rapid classification of large numbers of men.

In the following paragraphs a number of proposals in the field of selection and classification are described. Some of these proposals deal with suggested classification procedures which might be initiated and subjected to analysis and evaluation. Others outline problems of importance to the Navy and indicate some tentative approaches toward their solution.

WHAT ARE THE MINIMUM AND OPTIMUM REQUIREMENTS FOR ASSIGNMENTS TO EACH TYPE OF MILITARY DUTY? The problem of determining the minimum and optimum requirements for assignment to the various military duties can be studied by the standard techniques of correlating quantitative predictive measures with measures of success, determining cutting scores, computing statistics describing selection cost, and the like. Obviously a satisfactory cri-

terion of success is essential for any study of this type. (See Chapter XIX and later discussion in this chapter). In addition to the usual type of statistical study, it would probably be fruitful to make clinical studies of the individual cases where prediction failed. Such a technique might bring to light important hidden factors influencing success in a particular job. It might reveal job characteristics which were incompletely described in the preliminary job analyses and which are found to be crucial in determining failure or success. These findings would perhaps indicate what line of investigation should be followed in developing selection devices for the improvement of classification procedures.

COULD SELECTION AND CLASSIFICATION BE IMPROVED THROUGH THE USE OF TWO OR MORE BATTERIES OF TESTS IN MULTIPLE-STAGE SELECTION? The use of one set of tests for two such diverse purposes as separating school from non-school material and separating candidates for one type of school from candidates for another type of school is likely to result in tests which are not entirely suitable for either purpose. Furthermore, the administration of an extensive battery of tests to an unselected group of persons is probably very wasteful of testing and test scoring time. It might be possible to administer a greater variety of tests without increasing the total time spent in test administration by the device of using a short *primary* battery, to sort men into rough categories, and additional *secondary* test batteries for the purpose of making more specific assignments.

In the first selection stage, the primary battery for enlisted men might consist of only two tests, a verbal test and a mechanical test, constructed in such a way as to reduce their intercorrelation as much as possible. Scores on these two tests could be used (1) to separate school candidates from the non-school material, and (2) to classify the school candidates as to general type of school for which they should be considered. The scheme might work somewhat as follows: (a) men low on both verbal and mechanical tests would be assigned to general detail; (b) men who are high on verbal but low on mechanical tests would be considered as candidates for "clerical" types of schools (yeoman, signal, radio); (c) men who are high on mechanical but low on verbal tests would be considered as candidates for "mechanical" types of schools (gunner's mate, basic engineering, metalsmith) and (d) men who are high on *both* tests would be considered as candidates for the more highly technical schools (fire controlman, radio materiel).

The second stage would be to administer to the men in groups (b), (c), and (d) secondary batteries of tests designed especially to aid in making assignments to schools within each group. For example, the

men in the "clerical" group might be given tests measuring aptitude or ability in spelling, clerical work, radio code, and blinker signaling which would aid in distributing men to the schools in that group.

In such a program each man would be given only "appropriate" tests and the time required for testing and scoring would probably be less than that required in the present procedure of giving every basic battery test to every man. Tests designed for specific purposes might also be more valid than the general-purpose test and permit a greater use of individual abilities. A research program to investigate the feasibility of such a plan in classifying both officers and enlisted men might yield important results.

WHAT IS THE SMALLEST NUMBER OF TESTS THAT CAN BE USED EFFECTIVELY IN A BASIC TEST BATTERY? The answer to the question of how many tests should be included in a basic test battery might depend to some extent on the administrative procedures that are adopted for use in making assignments. (It would of course be sounder to base the administrative procedures upon the results of experimental study of single-stage versus multiple-stage testing). If the procedure were to permit only one testing period during which all test data must be obtained, a larger number of tests would be required than if it were possible to make tentative assignments or rough sortings on the basis of a basic test battery and further testing to assist in making final specific assignments.

The decision as to how many and what tests should make up a basic test battery used in a single-stage testing procedure should be made on the basis of two kinds of information: (1) the similarities between tests, as indicated by their intercorrelations, and (2) the differential validities of the tests. On the basis of the data reported in Chapter VI on the Basic Test Battery, and under the conditions which obtained during World War II, four instead of eight tests (exluding the radio code test) would probably suffice. The four tests might include one verbal, one quantitative, one mechanical, and one clerical test. These are the types which have been included in Forms 4 and 5 of the Basic Test Battery. Whether or not such a battery will prove as adequate as the longer battery formerly used is, of course, subject to experimental investigation.

FOR WHAT BILLETS, IF ANY, ARE SPECIAL TESTS NEEDED? The argument for a multiple-stage classification procedure such as was described earlier rests upon the assumption that tests designed specifically for predicting success in particular billets are more adequate than general tests used for predicting success in many billets and training assignments. This assumption needs to be put to experimental test. But there is already evidence that special tests are

needed for predicting success in certain duties. The Navy has recognized the need for and has developed special tests to predict success as sonarman, radioman, and radio technician at the enlisted level. At the officer level, special tests include those for sonar, tactical radar, and radio specialist billets. Further validation of these tests is necessary. As classification procedures are refined, research is needed to discover if there are other fields where prediction could be improved through the use of tests specifically designed for specialized purposes.

WOULD APTITUDE TESTS INVOLVING PERFORMANCE BE SUPERIOR TO PAPER-PENCIL APTITUDE TESTS FOR SOME PURPOSES? Studies in certain fields which involve mechanical work, such as is taught in gunner's mate, torpedoman, and basic engineering schools, indicate that the existing paper-pencil tests of mechanical aptitude predict success on written and identification achievement tests fairly well, but fail to predict success if performance tests, which are designed to reflect proficiency in disassembling, assembling, repairing, and adjusting pieces of mechanical equipment, are used as the measure of achievement. Since performance tests are possibly a more valid criterion of success in such schools than are the verbal tests, it might be valuable to be able to predict scores on such tests.

One approach to the problem would be to devise aptitude tests which themselves involve performance, since in this way such factors as manual dexterity can be included in the aptitude measure. The test might be broad enough in scope to include a measure of learning, understanding of directions, and how to disassemble and assemble a piece of equipment. Thus the aptitude test would in miniature resemble the learning situation existing in many schools.

Such a test for enlisted men might be given somewhat as follows:

Each man in the group would be seated at a bench on which is placed a fairly complex and unfamiliar piece of mechanical equipment, such as a breech mechanism from a 40 mm. gun. A sound film, similar to the training films in actual use would be shown in order to give standard instructions on how to dissassemble and assemble the equipment. Then at a signal each of the men would attempt to perform the job himself. The score on the test might simply be the time required to complete the job.

Many variations of the test might be tried experimentally, varying the complexity of the equipment, the amount of training given, and the methods of scoring. For example, it might be found that more valid scores would be obtained with two showings of the film and two trials, or that the time required for the third trial, without additional training, would give best results. Some such mechanical aptitude test which involves performance might be found to predict

actual proficiency in mechanical work considerably better than the paper-pencil tests. Explorations in this area are worthy of time and effort.

It must be pointed out, however, that the correlations among various performance tests used to measure achievement in a single school tend to be low. High specificity of manual skills has frequently been found in studies of performance tests, and this greatly increases the difficulty of building a performance test which will predict success in a variety of situations.

To What Extent Should Proficiency Tests be Used in Classification? Success in a given activity can, in general, be predicted more accurately from measures of past success in that same activity than from measures of aptitude or from inadequate records of amount and kind of experience. College success, for example, can usually be predicted more accurately from high school grades than from "scholastic aptitude" tests. A careful analysis of the Navy classification program would probably reveal neglected opportunities for predicting success in a given billet from records of proficiency in similar work or activities.

In an amphibious training base, for example, crews were trained to man LCVP's (Landing Craft, Vehicle-Personnel). Three groups of the crew members, deckhand, signalman, and coxswain, were given essentially the same training for the first four weeks, since it was desirable that any man be able, in event of a casualty, to take over the duties of any other man. After four weeks, training became more specialized, with men designated as coxswains spending relatively more time in boat operation, signalmen in learning semaphore, blinker, etc., and deckhands in learning seamanship. Before any training, men had been classified as coxswain, signalman, or deckhand on the basis of age, physique, vision, previous experience, and Basic Test Battery scores. In this situation it would have been entirely possible to postpone classification until after the first four weeks of training and to base classification on actual ability in signaling and boat handling as determined by proficiency measures.

In Chapter XIII it was pointed out that the Basic Test Battery, designed primarily for use in the classification of recruits, was possibly inappropriate for use in the selection of candidates for advanced enlisted training. It seems quite likely that certain types of proficiency tests might be superior to the Basic Test Battery in this phase of the classification program.

A number of performance tests designed to measure achievement in enlisted schools are already available and could easily be used in advanced classification centers. The radio-code receiving tests provide a good example. Proficiency tests might also be given on ship-

board and the scores entered in a man's Service Record to aid in making assignments in case of transfer from one ship to another. Such tests for shipboard use have been successfully developed for 20 mm. gun crews.

An exploration of the possibilities in developing and using measures of attained proficiency of both officers and enlisted personnel in predicting their subsequent success in the same or in a similar activity is certainly in order. Such a project might yield evidence that measures of proficiency following a given amount of training or experience are superior to measures of aptitude as predictors of success, and that the possibilities of using such proficiency measures should be fully exploited. The data presented in Chapter XX on the relationship between school success and shipboard performance support this statement.

WOULD MEASURES OF INTEREST AND PERSONAL ADJUSTMENT BE HELPFUL IN CLASSIFICATION? Over a period of years it has become increasingly evident to psychologists and personnel administrators that any individual's performance in his job may be greatly influenced by his adjustment to living and by his pattern of interests to date. Some exploratory study has been done on the criterion of adjustment, on instruments and techniques for predicting adequate adjustment, and upon measures of interest. Most of the results, however, have been of such a tentative nature as not to be directly applicable to the military situation.

During the war a number of personal inventories were developed and used in the Navy, mainly for the purpose of screening. On the basis of scores on such tests it was found possible to reduce considerably the number of men to be interviewed by psychiatrists. In other words, the use of the tests was negative rather than positive. Further development of such measuring devices might result in instruments which could be used not only to identify the men likely to be troublemakers and misfits in the service, but also to aid in selecting officers and men who might be especially effective in positions of leadership and responsibility. This is an important area which deserves a coordinated program of study.

No measures of interests were used in the Navy except on an experimental basis. Interviewers at classification centers tried to make assignments consistent with the expressed preferences of the officers and men, but many persons had very inadequate knowledge of the duties involved in their choices. There was, furthermore, an attempt to develop preferences through the use of lectures and films glamorizing certain types of duty. The use of interests in classification was at a fairly superficial level. The development and use of suitable techniques for measuring interests and attitudes and relating them to

classification of naval personnel might lead to fewer requests for transfers and to better adjustment of each man to his job.

WHAT PERSONAL HISTORY ITEMS ARE VALID FOR CLASSIFICATION? A considerable amount of personal history is routinely obtained and entered on the records of all Navy personnel. Sample items are previous military duty, amount and kind of education, vocational training, occupational experience, leisure time activities, sports, talent for public entertainment, and highest position of leadership (Chapters II and III). The classification interviewer or interviewing officer evaluates this information subjectively in making his initial recommendations, and the information is recorded for later use in making assignments aboard ship or at advanced classification centers.

The usefulness of these personal history items has never been fully demonstrated. For populations already selected on the basis of test scores, it has been shown that age and amount of civilian education have little correlation with school success or performance aboard ship. Type of education (technical, general, or vocational) has not been carefully investigated, and civilian occupation has been studied only superficially. A more careful investigation of these factors might reveal that the predictive value of personal history items could be increased through refinements in evaluating and recording them. It might be found, for example, that refinements could be introduced into the recording of occupational data so that the essential skills involved in the civilian job (administrative ability, tact, originality, proficiency in the use of specific techniques and tools) could be related more definitely to the Navy jobs.

CAN THE INTERVIEW CONTRIBUTE TO MORE EFFECTIVE CLASSIFICATION? In the early part of the war, classification was performed almost entirely on the basis of data obtained by testing and by questionnaire forms. The procedure was often criticised for its lack of the more personal type of evaluation obtainable through the use of interviews. The procedures were later revised, and a number of classification interviewers were trained and put to work interviewing each recruit and making recommendations as to his assignment. A similar procedure was followed in the case of officers. The question of whether the interview results in better classification has frequently arisen, both in military and civilian experience, but few studies have been made to obtain an answer.

One study concerned with the reliability of the interviewers' ratings of recruits was carried out. Another study dealt with the validity of the quality ratings assigned to each recruit by the interviewer (Chapter XII). The latter study suggests that interviewing actually detracted from the success of the procedure; according to

this study, assignment to a service school on the basis of test scores alone would result in a more successful classification than assignment on the basis of interviewers' recommendations. Additional studies are needed for both enlisted and officer programs.

It is possible that such studies would verify the finding that, as at present used, the interview is a relatively ineffective classification technique. Present knowledge suggests that interviewers tend to weight too heavily the items of information which they themselves obtain, at the expense of aptitude test scores. A standardized procedure, called a "point score" method, was developed for aiding the classification interviewer to evaluate the various factors involved in making recommendations concerning the assignment of recruits. In this procedure uniform weights are assigned to carefully selected and explicitly stated factors. The use of such methods is probably more reliable and objective than the ordinary interview evaluation. The validity of such a procedure depends entirely upon the care with which the factors are selected, defined, and weighted. The results of validity studies of personal history items should be used to select the interview factors to be used and to establish specific rules for their weighting. More rigid requirements might also be set up for the application of test scores in evaluation. Factors and weightings would need to be determined and verified for specific schools and for various billets.

The suggestions here given for improving the interview tend in the direction of making the interviewer follow definite rules in recommending assignments. If followed to its logical conclusion, this tendency might lead to dispensing with the interviewer and using machine methods entirely. If efforts to improve the interview followed a different course, the conclusion might be that better methods of selecting and training interviewers would improve classification without introducing the mechanization of procedures. The research has not as yet gone far enough to predict which approach would ultimately lead to greatest improvement.

Research on Techniques of Training

The object of military training is to produce as rapidly as possible officers and men well-qualified to carry out the duties to which they will be assigned. The object of a training research program should, therefore, be to provide essential information which will assist directors of training programs in determining what should be taught, what specific teaching techniques would be most effective in various types of training, how long a training program is required,

how instructors should be selected and trained, how the motivation of trainees can be improved, and how effective the training program has been.

During World War II little research on training, other than that directly concerned with measurement of achievement, was carried to completion in the Bureau of Naval Personnel. The work of the Billet Analysis Section provided extensive and valuable information for the writing of curricula, but little in the way of concrete research results on curriculum methods was available. Accordingly, the naval training programs were planned on the basis of the best guesses that could be made by experienced naval officers and civilian educators. Except for work on methods of teaching radio code, very little research was done on the techniques of teaching.

In planning a research program on the problems of training, it is very important to keep in mind the critical situation obtaining in emergencies. Solutions which may be adequate for peacetime operations sometimes prove totally unadapted to the needs when large numbers of men are being inducted and must be trained in the shortest possible time to man newly commissioned ships equipped with technical devices requiring expert maintenance and operation.

Curricula must be planned in terms of the latest technical developments and operating doctrine. Adequate provision must be made for the extensive retraining of experienced personnel as well as for the initial training of large numbers of recruits and inductees whose civilian experiences may bear more or less relationship to the duties for which they must be trained. Plans must be made for the allocation and efficient use in training installations of equipment which is being installed on ships. Methods of selecting and training naval personnel to serve as instructors and training officers in service schools must be developed. Training techniques and training aids must be devised and evaluated in terms of their comparative efficacy in bringing about the desired learning results. The training research program should attempt to provide information which will be fundamentally useful in planning the development of training programs for a rapidly expanding Navy.

The list of training research problems which follows could be multiplied indefinitely. Each question could be asked specifically for each training program that exists now or is likely to exist in the event of another emergency.

How Can the Effectiveness of Training be Evaluated? The development of adequate methods of measuring the results of training is the key to quality control of personnel and to all research on problems of improvement in training. Satisfactory measures of results of training are useful for other purposes. For instance, meas-

ures of individual proficiency may be useful in the advanced classification of personnel and as criterion instruments for evaluating selection methods. They also serve as valuable incentives in motivating both trainees and instructors.

The proficiency of trainees in the job for which they are trained is the ultimate criterion of the success of a training activity. (For discussion of characteristics of an acceptable criterion, see Chapter XIX). Unless adequate criteria of school success and of success in duty are available, the results of experimental studies on training problems may be quite misleading. For example, if essay tests were used as the criterion of success in gunnery school, statistically significant evidence might be found that training in expository writing is more essential than practical work in range estimation or loading. Ridiculous as this illustration is, parallels can be found in actual practice. The criteria of successful training must be as valid, and their measurement must be as reliable as it is possible to make them. Without stable valid criteria, the validity of techniques for measuring the outcomes of instruction and of studies on the effectiveness of training courses and methods must remain in question.

Considerable work has been done on the development of methods for measuring the effects of training. Methods so far devised include the use of tests of various kinds and of ratings both of performance and of products constructed during training. But this work has been restricted to relatively few of the training activities in the Navy. Similar measuring devices should be developed for use in areas which have not yet been studied, and new techniques of measurement should be evolved where necessary. A method of evaluating the success of a trainee should not only be valid and reliable, but it should be practical in the sense that large numbers of trainees can be tested in a reasonable amount of time. In addition, it should possess face-validity or pertinence; that is, the critical elements in the technique must be clearly related to the most significant objectives of the training program, and the relationship must be apparent to trainees and instructors. Not all of the present measures of proficiency satisfy these criteria.

One of the major problems in the improvement of evaluation of training is the development of performance measures of higher reliability than has yet been attained. For many of the Navy's training programs, skill in the choice and application of techniques is of greater significance than verbal knowledge about the processes involved. But evaluation of these "practical" factors is more difficult of accomplishment. The observation and evaluation of trainee performance on such tasks requires more skill on the part of the observer than does the administration and scoring of an objective-type

test. The selection of representative tasks to provide an adequate sampling of the needed skills is complicated by (1) the apparent specificity (or lack of transfer) of the skills which comprise a job or rating, and (2) the lack of instructor time to give individual attention to the observation of test performances. The result is ordinarily to resort to rating or the development of short tests, both of which fail to provide an adequate range of discrimination among trainees. Averaging these ratings or short test scores tends further to restrict the effective range of discrimination. Development of reliable techniques for appraising performance may involve the development of synthesized or integrated units in both teaching and testing rather than the type of analysis which has characterized written tests. Experimentation in this area is one of the programs which should be given immediate attention.

Once a satisfactory criterion for evaluating the success of a training program is available, the way is open to introduce a variety of studies on such problems of instruction as curriculum, method, and motivation. Probably separate studies will be required for each specific type of training, since these problems grow out of (1) the nature of the teaching objectives and content of the program, and (2) the preparation and needs of the trainee population.

How Can More Effective Curricula and Training Schedules be Planned? Basic to any study of curricula and schedules for specific schools is a consideration of the general pattern of training for the naval service. As indicated in Chapters IV and V, the pattern of training followed during the war provided for an initial period of indoctrination followed by elementary training in schools or by duty assignments in which it was assumed that in-service training would be carried on. Advanced and specialized training was provided in schools, usually after the trainees had had some sea experience. Operational training provided for the adaptation of skills to new equipment and the development of team efficiency. The pattern of training proposed for enlisted men in the peacetime Navy contemplates an additional type of school between the recruit training program and elementary training for naval ratings. These "primary" schools are designed to provide basic training for a group of somewhat related billets or ratings.

Since research on curricula and training schedules has many facets, a number of questions dealing with different aspects of the total problem will be raised in the discussion which follows.

In the long run, would training aboard ship be more efficient for certain billets than training ashore? During a period of rapid mobilization, shore training is necessary because ships are needed for combat. There is evidence that shore training is better than in-service

training in preparing men for their duties in some billets, and it is possible that in certain other types of billets shore training is ineffective or less effective than shipboard training. If this should prove to be the case, provision must be made for training ships or for modification of the shore training so that it will more nearly approximate shipboard conditions with respect to whatever factors are found most important. For example, in gunnery training the motion of the deck can be simulated by the use of a movable platform on which the training pieces are mounted. Among the specific problems in this area which require further investigation are the following: To what extent can the training for each billet or job be carried out in organized schools ashore? What skills can be developed only by experience in the billet afloat? For what schools should sea experience be prerequisite? To what extent is shorebased operational training effective? Should personnel be sent directly from recruit training or indoctrination school into technical training or would basic and elementary training for billets be better motivated and more effective if a period of sea duty were interspersed?

What is the optimum length of a training program? The length of specific school programs has varied greatly during the war. Radioman training was initially 20 weeks but was reduced to 16 weeks. Recruit training was varied from 3 to 12 weeks in length. The shortest period of training that is consistent with satisfactory performance needs to be determined for each training activity. The development of such optimal length programs depends upon both the determination of effective methods of instruction and the development of adequate measures of progress and of competence. Conceivably, programs of variable length, depending upon the previous training and aptitudes of trainees, may result.

How much specialization in training is desirable? In the elementary gunner's mate schools, men were trained in the disassembly, assembly, maintenance, and repair of various types of guns, ranging from small arms to the 5″/38 dual-purpose rifles. For the majority of the men, much of this training was wasted. Those assigned to small landing craft had no occasion to work on guns larger than the .50 caliber machine gun. Those assigned to large ships would ordinarily be assigned to only one type of gun, 20 mm., 40 mm., or a 5″/38. If the men could be selected and earmarked for assignment to a particular type of gun, it would be possible to shorten the training period and probably to increase the efficiency of the trainee with respect to the gun to which he is assigned.

On the other hand, it can be argued that the Fleet needs men with a broader background of training. Perhaps it is better to train men for several related billets permitting them to choose or be

assigned their specific jobs aboard ship. The advantages of the broader type of training for purposes of developing leadership are obvious.

It might be desirable to provide both general and specialized training activities, assigning men who show qualities of leadership to the more generalized program of instruction. It is also well to keep in mind that the needs of war were met only by an increasing degree of specialization both in the rating structure and in the training programs provided. Comparable problems in officer training require study.

How can the need for refresher training be determined, and how can such training be given most efficiently? A few studies were made during the war which indicate that skills once acquired often deteriorate even when men are assigned to duty involving use of those skills. For example, it was shown that radiomen on destroyers that had been in action for a considerable period of time lacked many of the important skills that are not used routinely but only in emergency battle situations. Proficiency in receiving radio code may drop significantly even in the time intervening between the completion of training and assignment to a billet unless provision is made for refresher drill. Evidence has been found that the German navy had set up provisions for periodical evaluation of the proficiency of its radiomen. Studies need to be made of the curves of forgetting, under different conditions and for various skills required in the Navy's billets. On the basis of the findings, routines could be established for periodic testing. Skills found to deteriorate most rapidly should be tested most frequently. Such periodic testing might serve both as an incentive to maintain the skills at a high level and as a means of knowing what personnel can be depended upon under operational conditions. The next step would be to set up training programs, preferably aboard ship, ashore if necessary, to provide refresher drill whenever needed. It may be found that in some cases a continuous program of practical drill is needed to maintain skills which are not used routinely, so that when a need arises for the use of the skill, the safety of the ship will not be jeopardized. Fire drills are routine even in shore establishments where the fire hazard is comparatively small, but the training of key personnel for similar exigencies is sometimes neglected.

Closely related to the need for refresher training is the need for training which will adapt previously learned skills to the requirements of new equipment or new operating doctrine. It is not so difficult to determine this need, since the equipment is installed or the doctrine promulgated at some definite time. There is, nevertheless, a need to determine what program of instruction will be most

effective in bringing about efficient use of the new equipment or operational procedure. During the war, training courses, some as short as three hours and others as long as several weeks, were set up for the purpose of training experienced personnel. The entire program of refresher training demands adequate study.

WHAT TRAINING METHODS AND DEVICES ARE BEST ADAPTED TO THE VARIOUS PROGRAMS? Training programs during the war placed much emphasis on lecture and recitation methods, despite the fact that the objective was usually the development of skills. As experience accumulated and as equipment became more plentiful more emphasis was placed on shop work and operations. While this development appears to be in the right direction, studies are needed to establish the proper balance between theory and practice, to determine the best methods of presentation for specific topics, and to find optimum proportions and sequences of lecture, demonstration, drill, study, and examination time. The methods to be used in teaching probably need to be studied in terms of specific training programs rather than as general methods of teaching. Methods must be adapted to the content of the course and to the abilities and interests of the trainees in the classes. It is possible that such specific studies of particular types of training might lead to development of methods which could be generalized for types of content and of populations, but such a conclusion cannot be assumed in planning research on teaching methods.

Training aids are synthetic devices designed either to substitute for more expensive equipment or equipment that is scarce and critically needed by the Fleet, or to provide more effective means of presenting principles and operating procedures than would otherwise be available. There have been instances in which the availability of a training aid has determined the content of instruction without due regard for the pertinence of the learning gained, through the use of the aid, to the objectives of the training program. The utility and effectiveness of training aids should be evaluated in terms of their contribution to efficiency in teaching the skills, understanding, and knowledge desired in a given training program.

HOW SHOULD INSTRUCTOR PERSONNEL BE SELECTED AND TRAINED? Research in civilian education as well as in the Navy has indicated a fairly close relationship between what students think of the quality of their training and what they think of the proficiency of their instructors. Despite this fact, insufficient attention has been given in the past to research on the problems of selecting and training military instructors. In part this is due to the fact that civilian educational research has not provided clearly defined techniques for identifying good teachers or of ascertaining their intellectual and

personal characteristics. This situation may be due to the tremendous variety of civilian teaching positions and the varied educational philosophies. Since Navy training programs are aimed at the development of rather specific knowledges and skills, it seems possible that research on the identification and description of good teachers may be more fruitful in the Navy than in civilian education.

In considering the problem of selecting and training instructor personnel, three important facts must be kept in mind. First, it should be remembered that every officer and petty officer is expected to assume responsibilities in connection with the in-service training program; hence there probably is a communality between good petty officer qualities and good teacher qualities. Second, since there is considerable variety in Navy training, it probably will not be possible to pick good teachers generally. It seems unlikely, for example, that a good recruit training instructor could be described in the same terms as an instructor for an electronic technicians' school. Third, the Navy will continue to rotate men between shore and shipboard duty, hence it will not be possible to select personnel for school instructional duty only. In other words, Navy instructors, both officers and enlisted, must be selected from among those eligible for shore duty, and personnel serving as instructors will eventually return to sea duty.

The three major research problems which arise in connection with the selection and training of military instructors are: (1) How can the skillful and unskillful teachers be identified? This in other words is the problem of the criterion. (2) What are the personal and background characteristics of the good teacher? Since competence in a specialized field is accepted as the basic factor for availability as an instructor, then the additional characteristics to be considered in determining selection requirements for instructors should be those which distinguish competent fire controlmen, for example, as good teachers of fire controlmen, competent gunnery officers as good teachers of gunnery, competent commanding officers as good commanders of training centers. (3) What teaching techniques should be taught in instructor training schools? Especially important are the questions of what teaching techniques should be taught petty officers who are already competent in their specialties and of how the effectiveness of instructor training programs should be evaluated.

Some data pertaining to these problems can be obtained from materials published in civilian educational journals, but in view of the emphasis which the Navy places upon training as opposed to education, the problems are sufficiently unique to warrant intensive study within the naval service.

How Can the Morale and Motivation of School Trainees be

IMPROVED? The principal factor in the maintenance of morale and a high degree of motivation in the Navy's schools during the war was the anticipation that there would be immediate need for trainees to apply their newly learned skills and techniques and knowledges in situations where consequences of failure would be extremely grave. Men knew that their own survival, that of their shipmates, and the success or failure of their missions might depend on their ability to do the right thing in the right way at the right time. Even under this compelling pressure, there were influences which made for unsatisfactory morale. Since the end of hostilities, these influences have grown in importance.

During the war a large number of morale studies were made by the Army and a few by the Navy. These studies indicate that the morale factors generally effective in the services were (1) satisfaction with the job, (2) belief in the mission, (3) a realistic appraisal of the job ahead, (4) confidence in the training and equipment, (5) pride in one's unit or organization, (6) belief that one's individual welfare was a matter of concern, (7) relations between officers and enlisted men, and (8) faith in the cause and in the future.

Morale among trainees is influenced by these factors in a way peculiar to the training situation. Men who had enlisted in the expectation of immediate sea duty and combat action were not satisfied to be put in schools. Sometimes the instruction and equipment did not warrant confidence, and trainees were likely to recognize inadequacy and superficiality more quickly than their instructors. Pride in organization is difficult to maintain in a training station, since there is a feeling that, however necessary training may be, the status of a trainee is inferior to that of the man on operational duty. Furthermore, trainees often feel that they are in a transient status; they look upon their school period as an interlude and do not develop a feeling of "belonging" in the training group. Command relationships are somewhat complex in school situations, and trainees frequently do not know to whom they can take their problems. Consequently they feel that no one is interested in them as individuals.

In addition to the general morale factors, morale among trainees is affected by problems which are in many ways similar to the problems of motivation in educational institutions outside of the service. These problems include the presence or development of interest in the subject matter, participation by the trainee in the activities of the learning process, knowledge of progress made, belief in the usefulness of the learning material, and other such elements which contribute to good teaching and efficient learning.

A continuing program of morale appraisal is needed. Trainee

populations turn over rapidly, and there are frequent changes in the personnel assigned to training school staffs. Care must be exercised in determining the causes of unsatisfactory morale. Different causal factors may result in the same expressed grievance. Remedial and preventive measures must relate to causes rather than to symptoms if they are to be effective.

Research on improvement of morale and motivation in the Navy schools should be directed toward the development and validation of (1) means of discovering unsatisfactory morale conditions before they emerge as overt symptoms, (2) techniques of identifying the causes of unsatisfactory morale, (3) remedial measures, and (4) preventive practices.

Some work has been done on the development of attitude and opinion surveys designed to discover causes of discontent in the service at large and to obtain an evaluation of training by former trainees who have served in billets for which the training was designed. It has been found that achievement tests given at appropriate intervals serve to keep the trainees aware of their progress and supply incentives for serious effort. Speed tests of performance have been particularly effective in producing competition among trainees. While these devices are helpful, they are far from adequate.

The development and use of attitude and opinion measures for both trainees and school staffs might prove helpful in determining the causes of low morale in training stations. The use of interest measures as a factor in the selection of men for assignment to school training might prove valuable in the maintenance of motivation. Experiments with values of team training and group responsibility might also prove useful.

It is likely that the morale problems in the schools of the peacetime Navy will differ considerably from those which were observed during the war. Efforts should be made, therefore, to study the possibility of adapting the findings and techniques of morale studies in industry and of motivation studies in education to the needs of the naval service.

Methodological Problems

What personnel research was accomplished during the war on the problems of training and of selection and classification was performed under the hampering restriction that it should interfere as little as possible with the personnel procedures already in use and the routines already established. Such considerations are perhaps reasonable and justifiable during the actual prosecution of a war, but in peacetime the way should be open for unrestricted research on a scale sufficiently broad to obtain definitive findings. It might

well be argued that one of the primary purposes of a military organization in peacetime is to study and perfect its own organization and procedures; this is essentially research. In the material that follows three significant methodological problems are discussed briefly.

CAN SATISFACTORY CRITERIA OF SUCCESS IN PERFORMING MILITARY DUTIES BE DEVELOPED? Evaluation of the techniques used in classifying and training men for military duties depends upon the existence of valid and reliable measures of the success with which men selected and trained in accordance with specified techniques enter upon and perform their duties. This fact has already been stressed (Chapter XIX), but it cannot be too strongly emphasized that unsatisfactory criteria of success may and sometimes do lead to research findings that are seriously misleading.

The validation studies reported in previous chapters have used as criteria school grades, achievement test scores, and rankings and ratings of various kinds. Are these measures acceptable as criteria? In many cases one is forced to conclude that they are not; pertinent phases of performance are omitted, or the weight given to significant factors is not in accordance with their real importance, or extraneous factors are introduced, or the measures lack consistency, or they fail to provide an adequate range and precision of discrimination.

The techniques of measurement used in the development of criteria will necessarily vary from one type of duty to another. Objective practical performance tests may be satisfactory in some instances. In others it may be necessary to devise rating methods and to embark on a program for the training of raters. New devices may be needed; for example, motion picture photography might be used in obtaining permanent records which could be reviewed by panels of expert judges or which could be used in the training of raters.

The most important single problem of method which needs to be solved is the development of satisfactory criteria of success. Work on the criterion should be undertaken at the very beginning of each test development or research project, and under no conditions should its development be postponed until the experimental selection and training methods have been devised.

HOW CAN THE DESIGN OF EXPERIMENTS BE IMPROVED? Aside from the development of acceptable criterion measures, the careful planning of the design of experiments is of the highest importance in the prosecution of a research program. In the validation work carried on in the Navy during the war, it was not feasible to establish adequate control groups to test the efficacy of the classification and training procedures which were experimentally developed and

applied. Accordingly, the findings on their effectiveness are primarily empirical rather than experimental. For example, in the validation studies on selection tests, data could be obtained only for groups that had already been selected by whatever procedures were in use at the time, making it necessary to apply mathematical corrections for curtailment in range of scores on the selection tests being studied. For experimental purposes it would be much more desirable to plan a series of experiments in which men are assigned to training activities on a random basis, so that the full range of talent could be represented. The relationships between selection measures and measures of achievement would then be more fully apparent, and one could state with greater confidence which factors are most important for success.

Experiments on training programs should be conducted with groups of men carefully matched with respect to the factors known to be important for success in school. Only when the variability with respect to initial ability is carefully controlled can the effectiveness of various types of training be compared. It might be desirable, for a period of time, to earmark a large proportion of recruits as experimental subjects. These men would be distributed to training and duty in various schools and operational activities ashore and afloat in accordance with the plans of the experiments in classification and training techniques. Their experiences in the Navy would be carefully controlled, recorded, equated, and compared. Obviously, such assignment and control procedures might detract from the immediate operational efficiency of the units to which the experimental personnel were assigned. On the other hand, such basic experimentation can be carried on only during periods of peace; the temporary loss of operational efficiency during peacetime should be a small price to pay for the development of procedures which can be relied upon to increase the speed and efficiency of mobilization and training when such speed and efficiency are extremely critical factors.

ARE THE FINDINGS OF PEACETIME RESEARCH APPLICABLE TO A WARTIME SITUATION? Since the real test of the peacetime research is whether or not it leads to more efficient classification and training in an emergency, it is highly important to know to what extent the findings of research conducted in peace can be applied to the solution of wartime personnel problems. Two principal factors appear to cast some doubt upon the applicability of peacetime personnel research to wartime problems: (1) differences between the characteristic abilities and aptitudes of the service personnel obtained under the two conditions, and (2) differences in the attitudes of men toward military service under the two conditions.

If a considerable proportion of the Navy's personnel continues to be drawn from the selective service program, there will be less difference between characteristic abilities and aptitudes of the peacetime and wartime populations but probably a greater difference between their attitudes toward military service. If, as seems more likely, the personnel of the peacetime Navy is obtained by voluntary enlistment, considerable differences may be found in such factors as age, previous occupational experience, amount of school training, and the like; but the attitude of the volunteer group toward military service would likely be more nearly comparable to the attitude prevailing during the war years.

One approach to the problem of determining the extent of applicability of the research findings would be to repeat some of the researches completed during the war. If similar results are obtained on the repeated studies, the presumption of applicability would be supported for more carefully controlled studies in the same areas. Where the results differ materially from those previously obtained, the direction and degree of difference may indicate a means of developing techniques of adaptation.

Another factor which may affect the validity and the applicability of research findings for some time to come derives from the fact that there are relatively few men in the 20 to 30 age group who have not had some previous military experience. The Navy should not lose sight of the fact that techniques for the adaptation of skills and knowledge acquired in civil pursuits to the requirements of service billets constitute an important factor in an efficient program of mobilization. The Navy's personnel research activities should not, therefore, be confined too closely to the study of its own requirements and practices. The techniques used and the research carried on in connection with selection and training in education and industry should be continuously observed, and their applicability to the needs of the Navy should be explored.

Solutions for the problems which have been outlined in this chapter will require research studies of two general types: (1) those of a basic character, such as development of techniques of measurement and analysis which can perhaps be best investigated in the laboratory situations which some universities and research organizations afford, and (2) those of an operational nature which could probably be solved by the application of present techniques, but which, owing to lack of facilities or of long-range planning, have not thus far been adequately studied. The latter should be undertaken either by or under the immediate sponsorship of the Navy itself.

Civilian research had succeeded, before the war, in developing valid and reliable measures of verbal, mathematical, and mechanical aptitudes. These subsequently proved their worth as techniques in military classification and selection. In certain other phases of aptitude testing and in the appraisal of personal qualities, basic research had not yet developed sufficiently effective techniques. When the necessary fundamental research on the measurement of personality factors and interests has been accomplished, it will perhaps be possible to use such measures more effectively in the military situation. The experience of World War I greatly stimulated research in the measurement of intelligence and aptitudes. A similar increase in research on problems of the measurement of personality factors may be expected to follow from the needs for such measures which were demonstrated in World War II.

The lack of a carefully planned systematic approach to the study of the Navy's personnel problems before and during World War II may be corrected by the establishment of the Research Activity of the Bureau of Naval Personnel (Chapter I). Techniques are already available for carrying out necessary research on many phases of the selection, classification, and training processes. As proved techniques for study of the problems of welfare, morale, personal adjustment, and other factors in the administration of personnel management are developed, these problems can be accorded the attention which they deserve.

This chapter does not attempt to lay down a program for the Research Activity. It must be assumed that competent personnel will develop their own program and will progressively adapt and develop the requisite techniques for the solution of the problems which will continue to arise. It cannot be emphasized too strongly, however, that adequate solution of the Navy's personnel problems requires the systematic planning and prosecution of a comprehensive research program. The aims of this program should be to develop and prove the techniques which will make the Navy most effective as a fighting force by enabling it to mobilize in a minimum time with the least possible waste of manpower.

APPENDICES

APPENDIX A-1

STAFF AND ORGANIZATION OF TEST AND RESEARCH SECTION

AUGUST 15, 1945

TEST AND RESEARCH SECTION

Lt. Comdr. R. N. Faulkner
Officer-in-Charge
Lt. D. B. Stuit [1]
Assistant Officer-in-Charge

Lt. L. C. Fowler [1]
Administrative Assistant
Dr. H. R. Haggerty
Editor

SELECTION TEST UNIT

Bond, G. L., Lt. Comdr.
Officer-in-Charge
[1] Bloom, R. F., Lt.
[1] Brundage, E. G., Lt.
Coffey, W. C., Lt. (jg)
[2] Cruikshank, R. M., Lt. (jg) (WR)
[1] Darling, W. C., Lt.
Owens, W. A., Lt.
Shafer, H. M., Lt. (jg)
[1] Wexler, M., Lt.
Williams, E. B., Lt.
[1] Zirkle, G. A., Lt. (jg)
Johnson, W. F., Lt.

ACHIEVEMENT TEST UNIT

Ryans, D. G., Lt.
Officer-in-Charge
[1] Batchelder, H. T., Lt.
Bechtoldt, H. P., Lt. (jg)
[1] Carstater, E. D., Lt.
Cooper, J. B., Lt.
Jackson, J. S., Lt. (jg)
McWilliams, A. R., Lt. Comdr.
[1] Monroe, A. E., Lt.
Porter, R. B., Lt.
Schneider, A. E., Lt. (jg)
Tiedeman, S. C., Lt.
Graver, H. A., Lt.
Van Dyke, V. B., Lt.
Yampolsky, M., Mrs.

RESEARCH UNIT

Lannholm, G. V., Lt.
Officer-in-Charge
Pace, C. R., Dr.
Civilian Head
[3] Curtis, J. F., Ens.
[1] Gibson, D. L., Lt. (jg)
Hausknecht, G., Mr.
Maucker, J. W., Lt.
Embree, R. B., Lt.
[2] *Porter, E., Lt. (jg)*

RADIO MATERIEL UNIT

Feder, D. D., Lt. Comdr.
Officer-in-Charge
Gettys, L. E., Ens. (WR)
[1] Lawrence, W. R., Lt.
[1] Woods, E. H., Lt.
Blau, H., Mrs.

Persons formerly associated with the Test and Research Section but transferred before August 15, 1945, are listed in italics.

[1] Since promoted to Lieutenant Commander.
[2] Since promoted to Lieutenant.
[3] Since promoted to Lieutenant (junior grade).

APPENDIX A-2

STAFF OF NDRC PROJECT N-106 AND RESEARCH DEPARTMENT OF COLLEGE ENTRANCE EXAMINATION BOARD[1]

Name	Title
John M. Stalnaker[2]	Associate Secretary, CEEB Contractor's Technical Representative, Proj. N-106
Henry Chauncey[3]	Associate Secretary, CEEB
Harold O. Gulliksen, Ph.D.	Project Director Contractor's Technical Representative, Proj. N-106
M. Duane Bown, Ph.D.	Research Associate
Herbert S. Conrad, Ph.D.	Research Associate
Thelburn L. Engle, Ph.D.	Research Associate
Nicholas A. Fattu, Ph.D.	Research Associate
Norman O. Frederiksen, Ph.D.	Research Associate
Charles M. Harsh, Ph.D.	Research Associate
Evan R. Keislar, Ph.D.	Research Associate
Luzelle D. Mays, M.A.	Research Assistant
Joseph Miller, Ph.D.	Research Associate
Donald A. Peterson, M.A.	Research Assistant
Irving Robbins, Ph.D.	Research Associate
George A. Satter, Ph.D.	Research Associate
Merrell E. Thompson, Ph.D.	Research Associate
Ledyard R. Tucker, Ph.D.	Head of Department of Analysis

[1] Only the names of those members of the College Entrance Examination Board who assisted directly in the Navy's aptitude and achievement testing program are listed. In addition to those listed above the following members of the College Entrance Examination Board and the Princeton University faculty gave assistance at various times: Samuel S. Wilks, Carl H. Wedell, Lynnette Birdsall Plumlee, Edith M. Huddleston, and James N. Hillhouse.

[2] Resigned June, 1945.

[3] Assumed duties June, 1945.

APPENDIX B

LIST OF SELECTION AND CLASSIFICATION TESTS, ADVANCEMENT EXAMINATIONS, ACHIEVEMENT EXAMINATIONS, AND TECHNICAL PERSONNEL AIDS DEVELOPED OR APPROVED FOR OFFICIAL USE BY THE TEST AND RESEARCH SECTION[1]

Selection and Classification Tests

OFFICER TESTS

1. General and Special Aptitude Tests	
Officer Qualification Test	Forms 1, 2, 3
Officer Qualification Test	Form 10
Officer Classification Test	Form X-1
Officer Classification Battery	
Part I. Verbal Reasoning Test	Form 2
Part II. Mechanical Comprehension Test	Form 2
Part III. Mathematics Test	Form 2
Part IV. Relative Movement Test	Form 2
Part V. Spatial Relations Test	Form 2
Officer Selection Test	Form I, Books 1, 2 Form II, Books 1, 2
Pre-Radar General Aptitude Test	Form 1
General Mathematics Test	Form 1
General Physics Test	Form 2
Combat Information Center Aptitude Test	Form II
Sonar Pitch Memory Test (Phonograph Recording) Two Records	Form 1
Relative Movement Test	Form 2
2. Inventories and Psychiatric Screening Devices	
Biographical-Preference Inventory	Form 1
Officer Personal Inventory	Form 1

ENLISTED PERSONNEL TESTS

1. Basic Test Battery	
General Classification Test	Forms 1, 2, 3
Reading Test	Forms 1, 2, 3
Arithmetical Reasoning Test	Forms 1, 2, 3
Mechanical Aptitude Test	Forms 1, 2, 3
Mechanical Knowledge Test	Forms 1, 2, 3
Clerical Aptitude Test	Forms 1, 2
Spelling Test	Forms 1, 2
Radio Code Test—Speed of Response	Form 2

[1] These tests are used in Navy programs and are not available for distribution.

General Classification Test	Fleet Edition Form 1
General Classification Test	Fleet Edition Form 1-S
Arithmetical Reasoning Test	Fleet Edition Form 1
Mechanical Aptitude Test	Fleet Edition Form 1
Electrical Knowledge Test	Fleet Edition Form 1
Mechanical Knowledge Test	Fleet Edition Form 1
Clerical Aptitude Test	Fleet Edition Form 1

2. Special Selection and Classification Tests

Enlisted Qualification Test (WR)	Form 4
English Test	Form 1
Non-Verbal Classification Test	Form 1
Literacy Test	Form X-1
Applicant Qualification Test	Forms 1, 2
Airplane Matching Test	Form 1
Mechanical Comprehension Test	Mark 3
Winchmen and Hatchmen Selection Test	Form X-1
Radio Technician Selection Test	Forms 6A, 7A, 8A, 9A

3. Inventories and Psychiatric Screening Devices

Billet Preference Record	Form X-1
Billet Qualifications Blank (Men)	Form X-2 (M)
Billet Qualifications Blank (Women)	Form X-2 (W)
Enlisted Personal Inventory	Form 1
Enlisted Personal Inventory	Form 2
Experience Comparison Index	Form X-1
Personal Check List	Form X-4
Personal Inventory	Form X-1 (W)
Previous Duty Check List	Form X-1
Social Judgments Test	Form X-1

Advancement Examinations

Fundamental Knowledge Required of all Men in the Navy	One Test
Seaman	Test for one pay grade
Coxswain	Test for one pay grade
Boatswain's Mate	Tests for three pay grades
Gunner's Mate	Tests for four pay grades
Turret Captain	Tests for two pay grades
Mineman	Tests for four pay grades
Torpedoman's Mate	Tests for four pay grades
Quartermaster	Tests for four pay grades
Signalman	Tests for four pay grades
Fire Controlman	Tests for four pay grades
Fire Controlman (O)	Tests for four pay grades
Yeoman	Tests for four pay grades
Storekeeper	Tests for four pay grades
Hospital Apprentice	Test for one pay grade

Pharmacist's Mate	Tests for four pay grades
Ship's Cook	Tests for three pay grades
Baker	Tests for three pay grades
Commissary Steward	Test for one pay grade
Radioman	Tests for four pay grades
Radio Technician	Tests for four pay grades
Radarman	Tests for four pay grades
Sonarman	Tests for four pay grades
Carpenter's Mate	Tests for four pay grades
Shipfitter	Tests for four pay grades
Machinist's Mate (S)	Tests for four pay grades
Fireman	Test for one pay grade
Machinist's Mate	Tests for four pay grades
Motor Machinist's Mate	Tests for four pay grades
Electrician's Mate	Tests for four pay grades
Water Tender	Tests for four pay grades
Boilermaker	Tests for four pay grades
Aviation Machinist's Mate	Tests for four pay grades
Aviation Machinist's Mate (C)	Tests for four pay grades
Aviation Machinist's Mate (F)	Tests for four pay grades
Aviation Machinist's Mate (H)	Tests for four pay grades
Aviation Machinist's Mate (I)	Tests for four pay grades
Aviation Machinist's Mate (P)	Tests for four pay grades
Aviation Electrician's Mate	Tests for four pay grades
Aviation Radioman	Tests for four pay grades
Aviation Radio Technician	Tests for four pay grades
Aviation Metalsmith	Tests for four pay grades
Aviation Ordnanceman	Tests for four pay grades
Aviation Fire Controlman (Aviation Ordnanceman—Bombsight)	Tests for four pay grades
Aviation Ordnanceman (T)	Tests for four pay grades
Parachute Rigger	Tests for four pay grades
Photographer's Mate	Tests for four pay grades

Achievement Examinations

OFFICER TRAINING PROGRAM EXAMINATIONS

1. U.S. Naval Reserve Midshipmen's Schools (Deck), Standardized Examinations

Seamanship and Communications	Form I (and Form I, Mod. 1)
Engineering and Damage Control	Form I (and Form I, Mod. 1)
Ordnance and Gunnery	Form I (and Form I, Mod. 1)
Navigation	Form I (and Form I, Mod. 1)

2. Other Officer Achievement Examinations

CIC Final Achievement Examination	Forms 1, 2
Pre-Radar Final Achievement Examination	Forms 1, 2, 3, 4

ENLISTED TRAINING PROGRAM EXAMINATIONS

1. Basic Engineering (Class P School)
   - Written Final Achievement Examination — Forms I, II, III, IV
   - Performance Tests
     - Auxiliaries
     - Boiler Maintenance
     - Evaporators
     - Hand Tools
     - Hand Tool Gages
     - Instruments
     - Internal Combustion Engines
     - Operation (Fire Room)
     - Piping
     - Pumps
     - Refrigeration
     - Ship Fitting
     - Tracing Pipe System
     - Turbines
     - Valves
2. Diesel (Motor Machinist's Mates—Class A School)
   - Written Final Achievement Examination — Forms I, II, III, IV
   - Performance Test (Diesel Engine Operation)
3. Electrical (Electrician's Mates—Class A School)
   - Written Final Achievement Examination — Forms I, II, III, IV, V, VI
   - Performance Tests
     - A-C Equipment and Interior Communications
     - Wiring
     - D-C Equipment
4. Fire Controlmen (Class A School)
   - Written Final Achievement Examination — Forms I, II
5. Gunner's Mates (Class A School)
   - Written Final Achievement Examination — Forms I, II, III, IV, V, VI
   - Performance and Identification Tests
     - Performance Tests (Final) — Forms I, II
     - Identification Tests (Final) — Form I
     - Performance (Unit) Tests
       - 5"/38 Performance Test
       - 40 mm. Performance Test
       - 20 mm. Performance Test
       - .50 Caliber Browning Machine Gun Performance Test
       - Small Arms Performance Test
     - Identification (Unit) Tests
       - 5"/38 Identification Test
       - 40 mm. Identification Test
       - 20 mm. Identification Test
       - .50 Caliber Browning Machine Gun Identification Test
       - Small Arms Identification Test

6. Gyro Compass (Class C-1 School)
   Written Final Achievement Examination — Forms I, II
7. Lookout
   Written Final Achievement Examination — Forms I, II
8. Quartermasters (Class A School)
   Written Final Achievement Examination — Forms I, II
9. Radio (Class A School)
   Written Final Achievement Examination — Forms I, II
   Performance Tests
     Radio Code Receiving Examination, Directions for Administering and Scoring
       Record No. 1. Directions to Trainees and Practice Tests
       Record No. 2. Message Tests
       Record No. 3. Plain Language Tests
       Record No. 4. Plain Language Tests
   Equipment Operation Tests
10. Recruit Training
    Written Final Achievement Examination — Forms I, II, III
11. Signal (Class A School)
    Written Final Achievement Examination — Forms I, II, III, IV
    Performance Tests
      Flashing Light Receiving
      Flashing Light Sending
      Semaphore Sending
      Semaphore Receiving
      Flag Hoist Spotting
      Flag Hoist "Bending On"
12. Special Training Program
    Reading Achievement Examination — Forms I, II
    Reading Classification Examination — Forms A, B
13. Storekeepers (Class A School)
    Written Final Achievement Examination — Forms I, II, III, IV, V
14. Torpedomen (Class A School)
    Written Final Achievement Examination — Forms, I, II, III, IV
    Performance Tests (Final) — Forms A, B, C, D
    Performance Tests (Period)
15. Yeomen (Class A School)
    Written Final Achievement Examination — Forms I, II, III, IV, V
    Performance Tests
      Directions for Administering and Scoring Typewriting — Forms I, II, III
      Directions for Administering and Scoring Shorthand — Forms I, II, III

16. Spelling
    Spelling Achievement Examination — Forms I, II
17. Telephone Talker
    Telephone Talker Final Achievement Examination — Forms I, II
    Check List for Correct Handling of Sound-Powered Telephone
18. Radio Technician Training
    Pre-Radio Materiel Achievement Examination
    Forms I, II, III, IV, V, VI
    EE and RM Final Achievement Examination
    Forms I, II, III, IV, V, VI
19. Radar Operators (Class P School)
    Written Final Achievement Examination — Group I, Forms A, B; Group II, Forms A, B; Group III, Forms A, B

Technical Personnel and Research Aids

A. TECHNICAL MANUALS

Constructing and Using Achievement Tests
(A Guide for Navy Instructors)
Insert A: Example for Naval Training School (Gunner's Mate)
Insert B: Example for Naval Training School (Electrical)

B. RATING SCALES

Teacher Trainee Rating Sheet (Radio Materiel)
Teacher Trainee Rating Summary Sheet (Radio Materiel)
Instructor Rating Scale (Radio Materiel)
Radio Technician Rating Scale (Radio Materiel) Form 1 (Revised)
Petty Officer Rating Scale (Radio Materiel) Form 1

C. INFORMATION SURVEYS

Information Survey 1 (questionnaire)
A Study of Enlisted Training
Information Survey 2 (questionnaire)
A Study of Educational Services
Information Survey 3 (questionnaire)
A Study of Amphibious Forces Personnel

APPENDIX C-1

LIST OF RESEARCH PROJECTS COMPLETED BY THE TEST AND RESEARCH SECTION[1]

Officer Personnel

STUDIES OF TESTS

1. Sex, regional, and part-score differences in Officer Qualification Test (Form 1) scores.
2. Summary of scores made on the Officer Classification Test by Reserve Midshipmen and Indoctrination School classes.

PREDICTION AND VALIDATION STUDIES

1. Prediction of success in Pre-Radar Training.
2. Analysis of Officer Qualification Test scores of successful armed guard officers.
3. Prediction of success of Tactical Radar Officers.
4. Prediction of success of Fighter Director Officers.
5. Prediction of success in a Naval Reserve Midshipmen's School.
6. Prediction of success of officers assigned to special aviation billets.
7. Validation of officer selection requirements for destroyer billets.
8. Validity of selection requirements for motor torpedo boat officers.
9. Validation of selection requirements for amphibious (LSM) officers.
10. Analysis of the relationship between scores on certain parts of the Officer Classification Test and mid-term examination grades of Reserve Midshipmen.
11. Validation of selection requirements for submarine officers.
12. Validation of selection requirements for Torpedo Data Computer operators.
13. Prediction of success in Construction Battalion Officer Training Program.
14. Factors associated with success of officers assigned to duty on assault transports.
15. Preliminary report on the validity of the United States Navy Officer Classification Test for predicting success in Pre-Radar Training.
16. Preliminary report on the validity of the United States Navy Officer Classification Test for predicting success in Diesel Engineering Training.
17. Validity of the Officer Classification Test in selection and placement of officers in Gunnery and Fire Control School.

[1] These studies are unpublished and are not available for distribution.

OTHER STUDIES

1. A study of the performance of junior officers aboard ship.

Enlisted Personnel

STUDIES OF TESTS

1. Studies of scoring accuracy on Basic Test Battery.
2. Trends in Basic Test Battery scores—for each Naval Training Center and for all Centers combined.
3. Comparison of Code Aptitude Test (Thurstone) and Radio Code Aptitude Test—Speed of Response.
4. Basic Test Battery score patterns of graduates of certain Class A schools.
5. Analysis of test score qualifications of service school enrollees at two Naval Training Centers.
6. The use of additive scores on the Basic Test Battery for selection purposes.
7. Determining the content of the Basic Test Battery.

PREDICTION AND VALIDATION STUDIES

1. Prediction of success in Naval Training Schools (Signalman and Quartermaster).
2. Validation of age and educational selection requirements for the classification of enlisted personnel.
3. Analysis of certain selection requirements for aerial gunners.
4. Development and validation of selection requirements for radarman strikers.
5. The use of the Non-Verbal Classification Test in the selection of Steward's Mates for submarine service.
6. Analysis of Ortho-Rater Scores of experienced gunners.
7. Relationship between qualifications of enlisted personnel and their performance in the Fleet.
8. Analysis of selection requirements for Advanced Service Schools.
9. Predictive efficiency of the Navy Basic Test Battery at Naval Training School (Gunner's Mates and Electrical Hydraulics).
10. Elementary Service School validity data on Forms 1 and 2 of Basic Test Battery.
11. The validity of the United States Navy Basic Test Battery in the selection and placement of enlisted men in a Class B school, Naval Training School (Fire Control—Advanced).
12. Prediction of success in Bomb Disposal School.
13. Development and validation of psychiatric screening devices for the identification of combat fatigue cases.

OTHER STUDIES

1. The USS New Jersey Classification Project.
2. Relationship between Selection Department recommendations and assignment of recruits at the end of Recruit Training.

3. Analysis of procedures for the selection of enlisted personnel for submarine service.
4. Estimated reliability of Specialist (C) Interviewers' recommendations.
5. Analysis of causes of attrition in Radio School.
6. Analysis of quota system operations for Class A schools.
7. Study of attitudes of enlisted men at the Amphibious Training Center, Pacific Fleet (5 reports).
   - Report No. 1. Fundamental Motivations.
   - Report No. 2. Prestige of Amphibious Duty.
   - Report No. 3. Attitude Toward Officers.
   - Report No. 4. Job Satisfaction.
   - Report No. 5. Opinion of Shore Training.
8. Navy Men Appraise Educational Services (a series of 9 reports on the results of Information Survey 2)
   - Report No. 1. Getting the news of the day.
   - Report No. 2. Interest in talks and discussions about the war.
   - Report No. 3. Newsmaps.
   - Report No. 4. Appraisal of the Navy's role in keeping men up-to-date on the news.
   - Report No. 5. Participation in USAFI.
   - Report No. 6. Participation in off-duty classes.
   - Report No. 7. Post-war educational plans.
   - Report No. 8. Accreditation services.
   - Report No. 9. Interest in war orientation readings.

APPENDIX C-2

TOPICAL LIST OF RESEARCHES BY NDRC PROJECT N-106, COLLEGE ENTRANCE EXAMINATION BOARD ON THE NAVY'S APTITUDE TESTING PROGRAM [1]

Officer Personnel

DEVELOPMENT AND VALIDATION OF APTITUDE TESTS

1. Preparation of the United States Navy Officer Qualification Test—Form 1. OSRD Report No. 1273. March 12, 1943.
2. Analysis of the N.R.O.T.C. Selective Examination (Form C) and suggestions for its revision. OSRD Report No. 1290. February 15, 1943.
3. Development and validity of the United States Navy Officer Qualification Test. OSRD Report No. 3186. January 7, 1944.
4. A statistical evaluation of the United States Navy Officer Qualification Test—Forms 2 and 3. OSRD Report No. 3978. August 4, 1944.

Enlisted Personnel

DEVELOPMENT AND VALIDATION OF APTITUDE TESTS

1. Naval Aptitude Tests: A. Reliability; B. Scoring Accuracy; C. Scoring Procedures. OSRD Report No. 1127. December 10, 1942.
2. Averages, standard deviations, and intercorrelations of Navy Aptitude Tests. OSRD Report No. 1536. June 7, 1943.
3. Factor analysis of the New United States Navy Basic Classification Test Battery. OSRD Report No. 3004. September 29, 1943.
4. Item analysis of Navy Aptitude Tests. OSRD Report No. 3039. December 30, 1943.
5. Validity of Navy Aptitude Tests in service schools at the Great Lakes Naval Training Station. OSRD Report No. 3245. January 31, 1944.
6. Statistical analysis of the Mechanical Knowledge Test. OSRD Report No. 3246. January 28, 1944.
7. Validity of an experimental battery of aptitude tests at the Ordnance and Gunnery Schools, Washington Navy Yard. OSRD Report No. 3619. April 29, 1944.
8. Selection of items for the U.S. Navy General Classification Test—Form 2 and the U.S. Navy Tests of Reading and Arithmetical Reasoning—Form 2. OSRD Report No. 3756. June 8, 1944.
9. The construction and vadidation of an Arithmetical Computation Test. OSRD Report No. 4556. January 8, 1945.
10. A further study of the validity of the Arithmetical Computation Test. OSRD Report No. 5302. July 3, 1945.

[1] These researches were conducted in collaboration with the Test and Research Section, Bureau of Naval Personnel. The titles of tests constructed in this program and published by the Bureau of Naval Personnel are listed in Appendix B.

DEVELOPMENT OF IMPROVED MEASURES OF ACHIEVEMENT

1. Memorandum on service school grades. OSRD Report No. 3177. January 6, 1944.
2. The development of achievement tests for Gunner's Mates Schools. OSRD Report No. 5259. June 25, 1945.
3. Achievement examinations for Signalman School. OSRD Report No. 5460. August 20, 1945.
4. Development of achievement tests for Class A Naval Training Schools (Torpedomen). OSRD Report No. 5520. August 31, 1945.
5. Development of achievement tests for the Landing Craft School, Coronado, California. OSRD Report No. 5670. September 14, 1945.
6. The development of performance tests for use in Class A Electrical Schools. OSRD Report No. 5666. September 13, 1945.
7. The development of Radio Code Receiving Examinations. Project Memorandum No. 18. September 27, 1945.

IMPROVEMENT OF RECORDS

1. A procedure for sorting the Enlisted Personnel Qualifications Cards. OSRD Report No. 4689. January 22, 1945.
2. Classification data available to ships' officers. OSRD Report No. 5145. May 30, 1945.
3. An experimental personnel record system for shipboard use. OSRD Report No. 5303. July 3, 1945. (Prepared in collaboration with the Enlisted Classification Section and the Test and Research Section, Bureau of Naval Personnel.)

SELECTION AND CLASSIFICATION PROCEDURES

1. An evaluation of the Personal Inventory for predicting success in Parachute School. OSRD Report No. 4870. March 28, 1945.
2. An evaluation of the Personal Inventory and certain other measures in the prediction of submarine officers' evaluations of enlisted men. OSRD Report No. 5557. September 7, 1945.
3. Predicting success in service school from the order of assignment. OSRD Report No. 5556. September 7, 1945.
4. The use of test scores and Quality-Classification Ratings in predicting success in Electrician's Mates School. OSRD Report No. 5667. September 13, 1945.

GENERAL METHODOLOGICAL REPORTS

1. Selection of test items by correlation with an external criterion, as applied to the Mechanical Comprehension Test—OQT O-2. OSRD Report No. 3187. January 8, 1944.
2. Characteristics and uses of item-analysis data. OSRD Report No. 4034. August 19, 1944.
3. The preparation of norms for the Fleet Edition of the General Classification Test. OSRD Report No. 4242. October 10, 1944.
4. A statistical evaluation of the Basic Classification Test Battery (Form I). OSRD Report No. 4636. May 14, 1945.

APPENDIX D-1

SAMPLE DIRECTIONS FOR ADMINISTERING A PERFORMANCE TEST

GENERAL DESCRIPTION OF THE 20 MM. PERFORMANCE TEST

I. *Purpose.* The purpose of this performance test is to measure the ability of students to perform certain tasks involved in the maintenance and operation of the 20 mm. gun.

II. *General procedure for administering.* With nine guns and three mounts, nine students may be tested at one time. About 25 minutes is required to test one group of nine students. This includes the time needed for giving directions, changing station, etc. Six runs would then be necessary to test a class of 50 students, and the total time would be about two hours and a half. Within this time a written test and other tests may also be administered.

The procedure for administering the performance test is as follows:
Nine working stations for students are prepared.

A. At three stations (Station A) students disassemble and assemble the trigger mechanism, the magazine interlock mechanism, and the double loading stop mechanism.

B. At three stations (Station B) the students, working at guns on mounts, are required to remove and replace the barrel, remove and replace the breech face piece, cock and uncock the gun, and place a magazine on the gun and remove it.

C. At three other stations (Station C) the students are required to disassemble and assemble the trigger casing group and the breechblock.

Thus nine students may be tested at one time.

One student and one proctor are placed at each of the nine stations. At a signal from the timekeeper, the proctor tells the student what to do, and at another signal the students begin work. Each proctor observes the work of the student at his station and marks a record form to indicate whether or not the specific operations were correctly performed.

In the meantime the timekeeper writes numbers on a blackboard indicating the number of half-minutes which have elapsed since the signal was given to begin work. When the student completes his job, the proctor copies the number from the blackboard in the space on the record sheet, indicating the time required by the student to complete the job. A maximum of five minutes is allowed.

When time is called, students stop work and the proctors get the gear in order for the next student. Then the students change stations and the procedure is repeated with each student working on the second of his three jobs. After time is called and the gear is again put in order, students again change stations, each going to the third of his three jobs. The test is scored in terms of time required and the number of errors made.

PRELIMINARY ARRANGEMENTS

I. *Assemble all necessary materials.* The materials needed for administering the performance test for the 20 mm. gun are as follows:

1. Six 20 mm. guns on tables with shoulder rests and barrel springs removed.
2. Three 20 mm. guns on mounts.
3. The necessary tools at each station.
4. A supply of record sheets. A sample record sheet is included in this booklet. At least one record sheet for each student to be tested must be mimeographed.

5. Copies of the *Directions for Proctors.* One copy of these directions is included in this booklet. Additional copies should be prepared so that each proctor may have a copy.
6. Directions for administering the test. A copy of these directions is included in this booklet.
7. A stop watch or a watch with a second-hand.
8. A blackboard, chalk, and eraser.
9. Pencils for the proctors.

II. *Conduct a training session for proctors.* One instructor is needed to serve as timekeeper, give the general directions, and supervise the test. One proctor is needed to observe and record the performance of each of the nine students. It is preferable to use instructors as proctors; but if enough instructors are not available, students in the last week of the Gunner's Mate course can do a satisfactory job, provided they are given adequate training. A training session of at least one hour should be conducted the day before the test is to be given. In training proctors it is necessary to give them instruction on the following points:

1. Give them a review of the gun and make sure they can speedily perform the tasks which they are to proctor.
2. Make sure that they understand the method of recording observations and time on the record sheets (see *Directions for Proctors*).
3. Make sure that they understand that they are *not to coach* the students during the test; proctors are to aid students only when necessary to prevent a student from injuring himself or damaging the equipment.

III. *Arrange the equipment in such a way that the test can be administered efficiently.* A suggested arrangement of equipment is shown in the accompanying diagram.

IV. *Make sure that the guns are properly assembled at the beginning of the test.* The guns on mounts should be completely assembled. Those on tables should be completely assembled except that the shoulder rests and the barrel springs are removed.

DIRECTIONS FOR THE TIMEKEEPER

I. Place one proctor at each gun.
II. Make sure that each proctor has a pencil.
III. Give each proctor one record sheet.
IV. Station one student at each gun.
V. Say: "WRITE YOUR NAME, CLASS NUMBER, AND BILLET NUMBER ON THE RECORD SHEET."

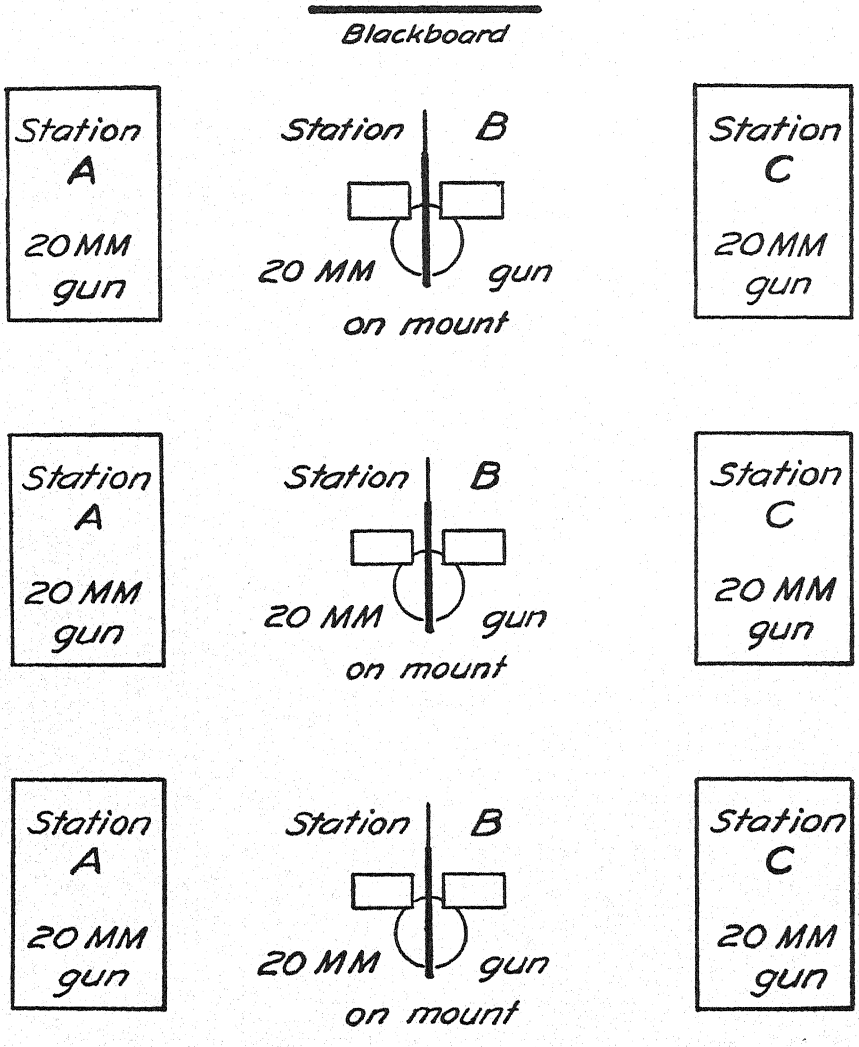
The proctors should have the students fill in the blanks on the record sheet. Then the proctors place the record sheets conveniently for recording their observations.

VI. Say: "THE PROCTORS WILL TELL YOU WHAT TO DO. DO NOT START WORK UNTIL I SAY BEGIN. WORK CAREFULLY. IF YOU KNOW YOUR JOB YOU WILL HAVE PLENTY OF TIME. DO NOT HURRY. YOU GET NO CREDIT FOR EXTRA SPEED. PROCTORS, GIVE DIRECTIONS." Wait 15 seconds, then say: "BEGIN."
VII. Immediately write *1* on the blackboard. After one-half minute erase the *1* and write *2*. After another half-minute erase the *2* and write *3*, and so on. At the end of five minutes say: "STOP." Erase the *10* and write *11*.
VIII. Say: "PROCTORS, MAKE SURE THAT THE RECORD SHEETS ARE CORRECTLY MARKED; THEN SIGN YOUR INITIALS." Allow time for the proctors to check the record sheets; then say: "PROCTORS, GET YOUR GEAR IN ORDER. STUDENTS STAY AT YOUR STATION UNTIL TOLD TO MOVE."
IX. When the gear is in order, say: "TAKE YOUR RECORD SHEET AND GO TO THE NEXT STATION." The proctors show the students where the next station is located.
X. Then say: "PROCTORS, GIVE DIRECTIONS." Wait 15 seconds and then say: "BEGIN."

XI. Proceed with steps VII to X, and repeat until each man has been at all three stations.

XII. Collect the record sheets and send the students back to the classroom.

XIII. Make sure the gear is in order for the next group of students. Repeat steps III to XII for each new group until the entire class has been tested.



Suggested Arrangement of Equipment for the 20 MM Performance Test

XIV. It is necessary to check the work of the proctors during the test to make sure they are not coaching the students and that they are recording their observations properly.

XV. Collect the copies of *Directions for Proctors* before discharging the proctors.

DIRECTIONS FOR PROCTORS

20 MM. PERFORMANCE TEST

You are to assist in administering a performance test to men studying the 20 mm. gun. You have two responsibilities:

1. Observe the performance of the student at your station and record it on his record sheet.
2. Keep the gear in order at your station.

Remember that you are helping to *give a test*. The student is on his own, and you must *not* interfere or give aid of any kind unless he is about to injure himself or damage the equipment.

I. At the beginning of the test you should have at your station:

   A. One 20 mm. gun. (At Station B, the gun must be on a mount.)
   B. The tools necessary to perform the job at your station.

II. When the instructor gives the order, have the student fill in his name, class number and billet number. Then take back the record sheet and put it in a convenient position for marking.

III. When the instructor tells you to "give directions" tell the student his first job. For example, if you are at Station B, say: "REMOVE THE BARREL FROM THE GUN."

IV. If the student performs a job correctly, circle *Yes* after the job on the record sheet; if not, circle *No*. When the student is finished with the first job, tell him to do the second, and so on. For each job, circle either *Yes* or *No* to indicate whether or not it was done correctly.

V. As soon as the student finishes the last job at your station, look up at the blackboard and notice the number written there by the timekeeper. Copy this number in the blank space under *TIME* at the right of the answer sheet. If the last job is not finished before the timekeeper says *stop*, write *11* in the blank space.

VI. When directed to do so by the timekeeper, get the gear in order for the next student.

VII. When the timekeeper tells the students to go to the next station, give the student his record sheet and tell him where his next station is.

VIII. Repeat steps II to VII until all students have been tested.

IX. Remember to give no aid or advice to the student unless necessary to prevent damage or injury.

RECORD SHEET

20 mm. PERFORMANCE TEST

Raw Score	Navy Grade

NAME.....CLASS.....BILLET NO.....

STATION A

Proctor's Initials			TIME
	1. Remove trigger plunger and spacer.	1. Yes No	
	2. Reassemble the trigger mechanism.	2. Yes No	
	3. Remove magazine interlock carrier spring.	3. Yes No	
	4. Reassemble magazine interlock mechanism.	4. Yes No	
	5. Remove the D. L. S. lever spring.	5. Yes No	
	6. Reassemble the D. L. S. mechanism.	6. Yes No	_____

STATION B

	7. Remove the barrel from the gun.	7. Yes	No
	8. Replace the barrel.	8. Yes	No
	9. Remove the breech face piece.	9. Yes	No
Proctor's	10. Replace the face piece.	10. Yes	No
Initials	11. Cock the gun.	11. Yes	No
	12. Uncock the gun.	12. Yes	No
	13. Place a magazine on the gun.	13. Yes	No
	14. Remove the magazine from the gun.	14. Yes	No

STATION C

	15. Remove parallelogram spring box.	15. Yes	No
	16. Remove parallelogram.	16. Yes	No
	17. Remove trigger hook & trigger hook holder.	17. Yes	No
Proctor's	18. Reassemble the whole group.	18. Yes	No
Initials	19. Remove hammer.	19. Yes	No
	20. Remove striker pin.	20. Yes	No
	21. Remove breech face piece.	21. Yes	No
	22. Reassemble the gun.	22. Yes	No

Number of No's circled ____ ____ Total Time

3 x number of No's circled ____

COORDINATION OF IDENTIFICATION, PERFORMANCE, AND WRITTEN TESTS

I. *Other tests.* The performance test is one of three tests which should be given each week. The other two tests are an identification test and a written test. The identification test which should be used is described in another booklet. The written test should be one prepared by the school. Suggestions which should be followed in preparing the written test may be found in the pamphlet, *Constructing and Using Achievement Tests—A Guide for Navy Instructors* (NavPers 16808).

II. It is possible to administer the identification test, the performance test, and the written test to a class of 50 men in a period of less than three hours. All examinations may then be administered on a Saturday morning. A scheme for coordinating the administration of these three examinations is illustrated in the diagram. *The procedure is as follows:*

1. Assemble the class in the room where the written examination is to be administered.
2. Divide the class into P-test groups, the size of each group being the number of men to whom the performance test can be administered at one time.
3. Send the *first* P-test group (for example, the *first ten men* on the muster list) to the room where the *performance* test is administered.
4. Send the *last* 25 men on the muster list to the room where the *identification* test is to be administered.
5. Start giving the *written* test to the remaining men.
6. As soon as the performance test is completed by the first group, this group is sent back to the classroom where they begin working on the written test. Then send the second P-test group (counting from the beginning of the muster list) to the performance test room. This procedure continues until all groups have been given the performance test.
7. As soon as the 25 men have completed the identification test, they should return to the classroom and begin work on the written test.
8. Approximately 45 minutes before it is anticipated that the performance test will be completed by all men, send the *first* 25 men on the muster list to take the identification test.

This procedure permits all the men to take all three tests within a period of three hours without conflict. Maximum use is made of the available testing time. It is necessary that all groups except the first and last group taking the performance test be interrupted once in their written test; this is not a serious objection if the written examination is of the short-answer or objective type.

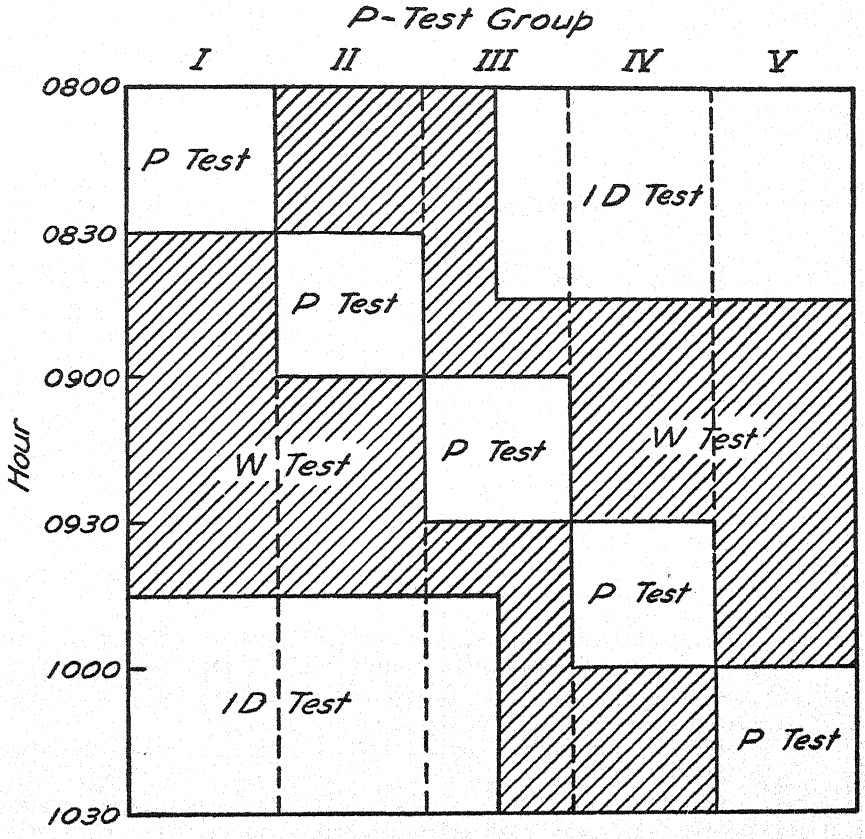


Diagram Illustrating the Method of Coordinating the Administration of Identification, Performance, and Written Tests

The diagram assumes that 10 men can be tested at once on the performance test. The system can be used just as easily with P-test groups of 8 or 9 men. In these cases, the last P-test group would contain fewer men than the other groups, if there are 50 men in the class.

Some care must be taken to prevent students from talking with each other about the written test items while going to and from the performance test.

APPENDIX D-2

SAMPLE DIRECTIONS FOR ADMINISTERING AN IDENTIFICATION TEST

DESCRIPTION OF TORPEDO IDENTIFICATION TEST

Purpose: To judge the trainee's familiarity with Mark 13 and Mark 15 torpedoes by using a wide sampling of disassembled parts from both torpedoes in order

(a) to test the trainee's ability to name the parts correctly, and
(b) to test his knowledge of how the parts function, or how improperly functioning parts should be adjusted or repaired.

General Procedure: Each torpedo part is laid on a large card on which is typed a list of names and a separate list of functions or adjustments. The cards are laid out about three feet apart along the sides of tables which are arranged in a continuous rectangular path around the room, so that the sequence of cards is easy to follow.

The students are provided with mimeographed answer blanks which they carry with them throughout the test. One student is stationed at each card. The test supervisor gives directions and each student answers the item at his station by writing two numbers on his answer sheet, one to indicate the correct name and one to indicate the correct function (or adjustment). The instructor then says, "Move," and each student moves to the next consecutively numbered station, where he examines the torpedo part and answers the item, as before. The procedure continues until each man has had a chance to answer all of the test items.

It is possible to test simultaneously as many men as there are parts involved in the test. With Forms C or D, 49 men can be tested in one hour and twenty-five minutes. With Forms A or B, 35 men can be tested in one hour. If only 30 men were tested on a 35-item test, the time required would be the same, but five stations would be vacant at any given time. If larger classes are to be tested, duplicate sets of parts and cards may be used to test several groups of trainees.

The advantages of this testing method over written or pictorial tests are that

(a) the man may pick up a part and manipulate it, to aid recognition or to aid in figuring out the function or adjustment; and
(b) the man is thus encouraged to be familiar with *actual* parts, rather than with manuals and pictures.

Security of Tests: The test supervisor should keep all materials in a locker while they are not in use. After the material is laid out for a test, the room should be secured until the time of the test. Instructors may tell students what types of examinations to expect, but should not reveal the content of specific test items.

PRELIMINARY ARRANGEMENTS FOR THE IDENTIFICATION TEST

I. *Assemble all necessary materials for one form of the test; these include:*

A set of ID Test cards. Each card bears a list of part names and a list of statements of function. These cards, with transparent covers, are supplied by the Standards and Curriculum Division of the Bureau of Naval Personnel (example in Fig. C).

A set of torpedo parts. The part-list for each of the four forms of the ID Test is included in this booklet (only Form A is shown here). Each list gives the name of the part to be placed on the correspondingly numbered card.

The parts need not be taken from service torpedoes. In many cases it will be satisfactory to use parts which have been discarded because of minor imperfections or excess wear. The parts should look fairly normal, however, and it is important that each part be from the proper torpedo (Mark 13 or Mark 15) if that is mentioned in the part-list.

Answer sheets. A sample answer sheet, usable for any form of the ID Test, is illustrated below. A large supply of these may be mimeographed and kept on hand.

A supply of extra pencils for students who lack pencils or pens.

A clock or watch with a second-hand.

II. *Arrange to have trained personnel available for administering the test.* One instructor is needed to supervise the test, and it is well to have one or two other instructors who are familiar with the test procedure to assist in proctoring, especially at the beginning of the test.

III. *Lay out the test materials an hour before the test is given.*

Arrange tables in the testing room to form a fairly continuous route around the room. *Plan on at least three feet of table space for each test item.* In a large room a rectangle of tables may suffice. Or a variation such as shown in Figure A, will care for a 35-item test in an ordinary room. This permits the proctors to stand inside the enclosure and observe the students, who are working around the outside of the table enclosure.

For a 49-item test, if space is limited, it may be necessary to have students work on both sides of the tables, and to arrange the tables in some form such as illustrated in Figure B.

Lay out the ID Test cards on the tables. The cards should be at least three feet apart, and cards from the first half of the test are alternated with cards from the last half of the test, as illustrated in Figure A or B. Thus consecutive numbers are two stations apart, so that a student can not easily see the item he last completed or the next item to which he will move. The students move two stations to the right each time. Thus, to respond to all items, a student must make two complete rounds of the tables.

Lay each torpedo part on the upper half of the corresponding card. Place one answer blank under the edge of each card (to prevent the blanks from blowing away).

IV. *Prepare a scoring key and check the materials.* An instructor should take the entire test, marking his answers with red pencil on a regular answer blank. Check these answers against the list of correct answers for whichever form of the test is being used. This is a check that each part is placed on the right card, and any errors can be corrected before the test begins. The answer sheet marked in red is kept for use as a scoring key.

ADMINISTERING THE TEST

Station one man at each card, and provide pencils for men who do not have pencils. Hold up an answer sheet where all men can see it, point to top line and say:

"WRITE YOUR NAME, CLASS NUMBER, AND BILLET NUMBER IN THE SPACES AT THE TOP OF YOUR ANSWER SHEET."

When all have finished, say:

"THE CARD IN FRONT OF YOU HAS A NUMBER IN THE UPPER LEFT-HAND CORNER. FIND THE CORRESPONDING CARD NUMBER ON YOUR ANSWER SHEET AND CIRCLE IT, TO SHOW YOU WHERE TO BEGIN MARKING YOUR ANSWERS." (pause)

"ON EACH CARD IS A TORPEDO PART. YOU MAY PICK IT UP AND EXAMINE IT IF YOU WISH. BELOW IT ON THE CARD IS A LIST OF *NAMES,* ONE OF WHICH IS CORRECT. SELECT THE CORRECT NAME, NOTICE THE NUMBER IN FRONT OF IT, AND WRITE THIS NUMBER IN THE '*NAME*' BLANK FOLLOWING THE CARD NUMBER THAT YOU CIRCLED." (Hold up an answer sheet and point to the 'Name' column.) "DO THIS NOW." (Allow about one-half minute.)

"NOW READ THE LIST OF *FUNCTIONS* ON THE CARD. SOME OF THESE ITEMS REFER TO ADJUSTMENT OR REPAIR OF THE PARTS, BUT IN ANY CASE DECIDE WHICH STATEMENT IS *MOST CORRECT* FOR THE

Class..... Billet No..... Name.....

TORPEDO IDENTIFICATION TEST Score

IN THE FIRST SPACE AFTER THE CARD NUMBER, WRITE IN THE NUMBER CORRESPONDING TO THE CORRECT NAME OF THE PART. IN THE SECOND SPACE, WRITE IN THE NUMBER CORRESPONDING TO THE FUNCTION THAT SEEMS TO YOU MOST CORRECT.

Card No.	Name	Function	Card No.	Name	Function
1			26		
2			27		
3			28		
4			29		
5			30		
6			31		
7			32		
8			33		
9			34		
10			35		
11			36		
12			37		
13			38		
14			39		
15			40		
16			41		
17			42		
18			43		
19			44		
20			45		
21			46		
22			47		
23			48		
24			49		
25					

Sample Answer Sheet

PART AT YOUR STATION, AND WRITE THE NUMBER OF THAT STATEMENT IN THE *SECOND BLANK* FOLLOWING THE CARD NUMBER. DO THIS NOW."

Allow two or three minutes while the proctors check to make sure that the students are recording answers in the correct spaces. Do not tell a student whether or not his answer is correct. Then say:

"NOW YOU SHOULD HAVE WRITTEN A NUMBER IN EACH OF THE SPACES FOLLOWING THE CARD NUMBER THAT IS CIRCLED ON YOUR ANSWER SHEET. YOU ARE TO PUT A NUMBER IN EVERY BLANK AS YOU CONTINUE THE TEST. IF YOU AREN'T SURE, MAKE THE BEST GUESS YOU CAN."

"YOU ARE TO TAKE THE ANSWER SHEET WITH YOU WHEN YOU MOVE, BUT LEAVE THE TORPEDO PART ON THE CARD WHERE YOU FOUND IT. WHEN I TELL YOU TO MOVE YOU WILL GO *TWO* STATIONS TO YOUR RIGHT TO FIND THE NEXT CONSECUTIVELY NUMBERED CARD. DECIDE ON THE NAME AND FUNCTION OF THE PART AT THAT STATION, AND BE SURE TO WRITE YOUR ANSWERS IN THE SPACES AFTER THE NUMBER OF THE CARD AT WHICH YOU ARE WORKING. YOU ARE NOT TO CIRCLE ANY MORE CARD NUMBERS. DO YOU UNDERSTAND THE DIRECTIONS?"

Answer any questions.

"AFTER YOU FINISH WITH EACH CARD, TURN YOUR ANSWER SHEET FACE DOWN AND WAIT UNTIL YOU ARE TOLD TO MOVE. NOW *MOVE* TO THE *SECOND* STATION TO YOUR RIGHT AND BEGIN WORK IMMEDIATELY."

At intervals of 1½ minutes say, "MOVE," until the men have completed the test. This interval may be varied to suit the speed of the students, in order not to rush them unduly, but do not bore them with waiting for a very slow man.

The proctors should continue to move about the tables to make sure that the students are recording answers in the proper places, and to detect or discourage cheating.

At the end of the test, collect the answer sheets and send the students back to their classroom. If another group is to be tested, check to see that all material is in order. When all groups have been tested, collect the materials and secure them in a locker until they are to be used again.

Plan to use a different form of the Identification Test on the next class. The forms may be used interchangeably, in random order.

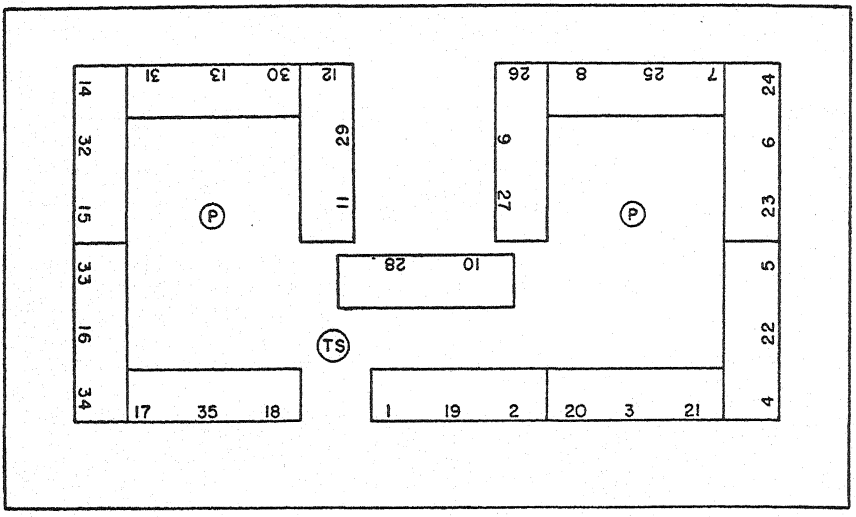


Figure A. SAMPLE ARRANGEMENT FOR 35-ITEM ID TEST

(TS) = TEST SUPERVISOR; (P) = PROCTOR; NUMBERS = ID TEST CARDS

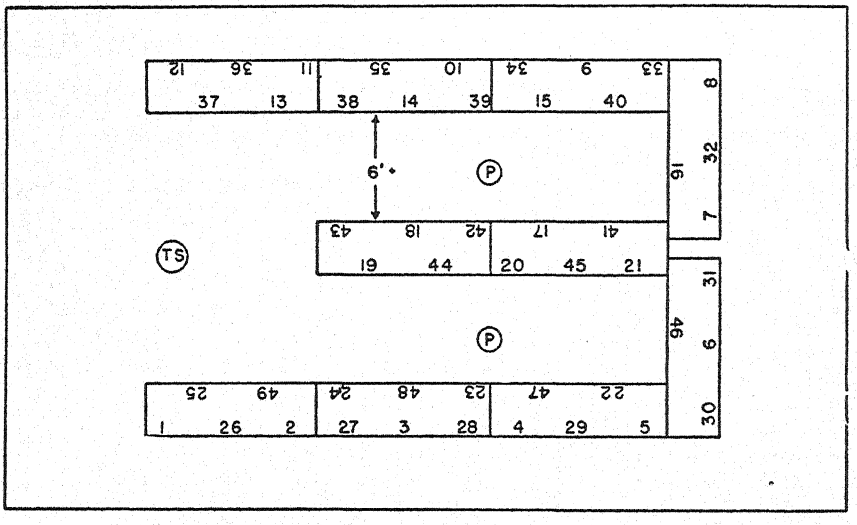


Figure B. SAMPLE ARRANGEMENT FOR 49-ITEM ID TEST

(TS) = TEST SUPERVISOR; (P) = PROCTOR; NUMBERS = ID TEST CARDS

TORPEDO IDENTIFICATION TEST, FORM A

	Torpedo Part	Correct Answer Name	Function
1.	Stop valve and carrier	5	5
2.	Piston fork	1	3
3.	Centering pin, spinning mechanism	4	3
4.	Stop valve plug and spindle, Mk. 13	4	1
5.	Delivery check valve guide	3	4
6.	Upper oiling bolt, Mk. 15	1	5
7.	Depth engine valve	4	2
8.	Starting piston body	4	5
9.	Mid-bearing outer race, Mk. 15	2	3
10.	Arming screw	3	4
11.	Oil pump shaft, Mk. 15	1	5
12.	Left pallet pawl	4	4
13.	Crosshead bushing, Mk. 15	3	4
14.	Air release mechanism spring	3	5
15.	Delivery check valve	2	4
16.	Four-part spindle bushing, Mk. 13	3	4
17.	Air strainer (short)	2	4
18.	Spring bell crank	4	5
19.	Stop valve plug and spindle, Mk. 15	4	4
20.	Exploder worm wheel	3	5
21.	Self-aligning bearing race, Mk. 15	4	2
22.	Reducing valve, Mk. 15	3	4
23.	Oil pump connecting rod, Mk. 15	3	2
24.	Vertical steering engine	2	4
25.	Main drive gear bushing, Mk. 15	2	1
26.	Valve bell crank	3	1
27.	Exploder worm wheel	4	3
28.	Charging valve bushing	2	4
29.	Unlocking rack	2	5
30.	Gyro wheel bearing, locking end	2	4
31.	Air check valve guide plug	4	1
32.	Top commercial bearing, Mk. 15	2	4
33.	Charging check valve	3	3
34.	Valve connection rod, Mk. 15	4	5
35.	Pendulum tool, 414A	2	4

SCORING THE TEST

The scoring key prepared before the test on a regular answer sheet has the correct answers written in red pencil.

Fold the scoring key along the right-hand edge of the first "Name" column; place the folded edge next to the "Name" column of the paper which is to be scored; and compare the answers. Make a heavy red pencil mark through every wrong answer on the student's paper. An omission counts as an error.

Fold the scoring key along the right-hand side of the first "Function" column; place this folded edge next to the corresponding column of the paper being scored; and again compare answers. Mark errors and ommissions with a red pencil.

Similarly fold the scoring key along the right-hand side of the "Name" column and of the "Function" column for cards 26 to 49, and score these columns in the same way.

Count the total number of red error marks in all four columns of a student's answer sheet. Subtract this number from the possible perfect score to get the number of correct answers. Enter this number of correct answers in the score box under the student's name at the top of the sheet.

When these scores are recorded in the class book or in the office records they should be converted to Navy Grades.

APPENDIX E-1

SELECTION OF ITEMS FOR THE OFFICER QUALIFICATION TEST, FORMS 1, 2, 3

The first step in the development of the tests was to devise an experimental form which contained an excess of items over the number needed. On the basis of item analyses, those items which failed to conform to the criteria adopted for inclusion in the tests could be discarded. The item analyses made possible an examination of faulty items with a view to a revision intended to eliminate the flaws in these items.

Two experimental versions of the Officer Qualification Test were administered, one to a sample of officer candidates, and the other to two groups of indoctrinees. Item analyses were then performed. On the basis of the first of these the items of Officer Qualification Test—Form 1 were selected, and from the second analysis items were chosen for Forms 2 and 3.

Data Yielded by the Item Analyses

BISERIAL CORRELATION BETWEEN AN ITEM AND THE SUBTEST IN WHICH THE ITEM OCCURS. This statistic, r_{bis}, indicates the degree to which an item measures the same thing as the subtest. The biserial coefficient of correlation is used as an index of the homogeneity of a test and of the discriminative power of an item—that is, the extent to which it discriminates between the high-scoring and the low-scoring testees. It is desirable, therefore, to select items which correlate positively and as high as possible with the subtest score.

ITEM DIFFICULTY. The statistic used as an index of the difficulty of an item was the proportion, p, of those attempting an item who answered it correctly. While this proportion shows the *ease* of the item rather than the difficulty, which would be $1-p$, the arithmetical work of subtracting p from a constant was considered unnecessary. In general, it is desirable to eliminate the items whose p-values are less than .05 (too difficult) or more than .95 (too easy). Biserial coefficients of correlation for items in these categories of difficulty are very unreliable.

It is considered desirable to have the bulk of the items of a test answered correctly by 40 to 80 per cent of the group, with the greatest concentration of p-values in the neighborhood of the mid-point between the chance score and 100 per cent.

THE NUMBER ATTEMPTING AN ITEM. This statistic, N_t, is the base number on which p is calculated. A person is judged to have attempted an item if he recorded an answer either to that item or to any subsequent item in the subtest of which the item is a part. This definition of N_t is based on the assumption that the person works systematically through the subtest, mentally attempting all items up to and including the last one for which an answer is recorded.

The criteria adopted for the selection of items for any of the three forms of the Officer Qualification Test depended on the distributions of p and r_{bis} for the items in the experimental forms, and on the excess of items, which determined the number which could be discarded. The general procedure followed was to discard as many as possible of the items of low biserial correlation. The exact minimal value of acceptable r_{bis} depended in each case on the number of items which could be discarded and on the distribution of values of r_{bis}.

Significance of Changes in Difficulty and Biserial Correlation of Items

When the difficulty (p) of an item and the biserial correlation (r_{bis}) of an item with its subtest were evaluated on two samples and tests were made to determine whether the item differed in either respect on the two administrations, the following criteria were adopted:

DIFFICULTY. In order to test the significance of a discrepancy in the p-value of an item, the *chi* square test was used. The degree of significance of a discrepancy in p as measured by *chi* square depends on three factors. The first factor is the magnitude of the discrepancy. The second factor is the number of cases. As N becomes smaller, the minimum p discrepancy which is significant increases. The third factor is the difficulty level of the item. The further the value of p is from .50, the smaller is the minimum discrepancy which is significant. This effect is most marked from .10 to .00 and from .90 to 1.00, but is also noticeable from .10 to .20 and from .80 to .90. Between .20 and .80 the level of difficulty has relatively little effect on the significance of a discrepancy.

A difference in p-value for the two administrations of an item was judged to be significant when the *chi* square test indicated that such a difference would be expected to occur by chance in less than one out of a hundred times.

ITEM-SUBTEST CORRELATION. In order to test the significance of a discrepancy between values of r_{bis} obtained from two administrations of the same item, the ratio of the discrepancy to its standard error was calculated. While it would be desirable to apply a test of the null hypothesis to these values of r_{bis}, there is no precise test of this hypothesis available for the difference between biserial correlation coefficients. The method used in the present studies is probably a fair approximation when the number of cases is large ($N = 500$). A difficulty is that in the estimate of the standard error the assumption of a normal distribution of r_{bis} is more or less erroneous.

The formula used in the estimation of the significance of a change in r_{bis} was:

$$\text{Significant Ratio} = \frac{r_{bis_1} - r_{bis_2}}{\sigma_{(r_{bis_1} - r_{bis_2})}}$$

where:

$$\sigma_{(r_{bis_1} - r_{bis_2})} = \sqrt{\sigma^2_{r_{bis_1}} - \sigma^2_{r_{bis_2}}}$$

In these formulae, subscript 1 denotes a statistic for the first group and subscript 2 denotes a statistic for the second group. The standard error of r_{bis} was calculated by the following formula:

$$\sigma_{r_{bis}} = \frac{\frac{\sqrt{pq}}{z} - r^2_{bis}}{\sqrt{N}}$$

where:

r_{bis} = the biserial coefficient of correlation.
p = proportion of persons answering the item correctly (based on N_t).
q = 1 — p.
z = the ordinate of the normal distribution curve corresponding to areas p and q under the curve.
N = the total number of cases.

APPENDIX E-2

DIRECTIONS FOR CONVERSION OF RAW SCORES ON ACHIEVEMENT TESTS TO GRADES IN THE NAVY 0-99 POINT SCALE

All elementary enlisted schools were directed during the war to assign grades on a scale which ranges from 0 to 99, with 63 as the lowest passing grade. Since the total possible raw score varied from test to test some convenient means was needed for converting raw scores to the Navy grading system. Likewise, some means was needed for converting total points earned by a trainee into a final course grade. Instructions issued to school staffs for making this conversion are described below.[1] The standardized achievement examinations constructed by the Test and Research Section were provided with the necessary conversion tables before they were distributed for use by the schools.

1. *First make a distribution showing the frequency with which the various scores on the test occur.* To do this, make a column of numbers ranging from 0 to the maximum score which can be obtained. This includes all the possible scores on the test. Then go through the test papers and place a tally mark opposite the number corresponding to each score which was obtained. These marks indicate the number of times each possible score actually occurred.

2. *Decide what Navy grade should correspond to the best score obtained on the test.* This is a matter of judgment on the part of the instructor and should be based on the frequency with which high scores are obtained, the instructor's judgment of the difficulty of the test, and the standards of achievement required. A paper with no errors need not necessarily be given a Navy grade of 99; it might be judged to be worth 95 or even less. It must be remembered, however, that the range of final grades obtained by averaging several different tests will be *less* than the range obtained on any one test. The reason for this is that it is unlikely that a man who makes an extremely high or low grade on one test will also make extremely high or low grades on all of the others. Therefore the Navy grade assigned to the best student on a single test must be *higher* than the highest grade desired on the final average grade.

3. *Decide what score on the test should correspond with a just-passing Navy grade (63).* This again is a matter of judgment. The instructor must take into account the frequency distribution which has been obtained, the proportion of students expected to fail, the difficulty of the test, and the standards of attainment required. Again it must be remembered that the range of *averaged* grades will be *less* than the range of grades on one test. Therefore the score corresponding to a Navy grade of 63 must be chosen in such a way that more students will fail the particular test than it is desired to have fail on the final average.

[1] This material is adapted from *Constructing and Using Achievement Tests,* Navpers 16808, Bureau of Naval Personnel, 1944.

4. *Prepare a translation graph.* The preparation of such graph is illustrated in the accompanying diagram. The following steps should be carried out in constructing a graph for translating raw scores into Navy grades:

a. On a piece of graph paper, place numbers from 0 to 99 along a line drawn at the bottom of the graph. Use one unit on the graph paper to represent one point on the Navy scale of grades.
b. On a vertical line drawn through the 0 on the base line, place the full range of raw scores which may be obtained on the test. Let each unit on the graph paper represent one score point. (If the number of raw score points is large, it may be necessary to let each unit on the graph represent two raw score points; if the total possible raw score is small, one raw score point may be represented by several spaces on the graph.) Make a raw score of 0 coincide with the 0 on the base line, and put the highest raw score obtainable at the top.
c. Make a dot on the graph which represents the intersection of a horizontal line extended from the highest raw score obtained in the test and a vertical line extended from the Navy grade which has been judged to be the equivalent of the highest score. This is point A on the graph—65 being the highest raw score in the test and 99 being judged its Navy grade equivalent.
d. Make a dot on the graph which represents the intersection of a horizontal line extended from the lowest raw score obtained in the test judged to be of passing quality and a vertical line extended from the Navy grade of 63. This is point B on the graph—25 being the lowest raw score judged to be of passing quality and 63 being the lowest passing grade in the Navy system.
e. Draw a straight line through the two points which have been located, and extend it in both directions to the edges of the graph. This is the "translation line."

5. *Translate each student's raw score to its equivalent Navy grade.* To do this, locate the raw score on the vertical scale at the left. Follow a horizontal line across until you reach the translation line. From that point on the translation line follow a line down until it intersects the base line. This point on the base line represents the Navy grade. Two examples will illustrate the procedure. (Figure 1-E-2). In Example A, a horizontal line from a raw score of 55 (Point C) intersects the translation line at D; a vertical line from D intersects the base line at 90 (Point E). In Example B, a horizontal line from a raw score of 37 (Point F) intersects the translation line at G; a vertical line from G intersects the base line at 74 (Point H).

6. *Set up a permanent translation graph.* The first time the test is given it will, of course, be necessary to use the data from only one class in making a translation graph. As soon as scores are available for 200 students, a frequency distribution of the grades obtained by the 200 students should be made and a translation graph prepared based on this distribution. This may be considered a permanent translation graph and should be

used for subsequent classes. However, if for any reason the general level of achievement of the students changes as indicated by scores on the test, it may be necessary to make a new translation graph, basing it on a frequency distribution obtained from a new sample of at least 200 students. For example, if all students consistently obtain Navy grades above 75 for a period of several weeks, it will be necessary to make a new translation graph; if this is not done the Navy grades assigned will be seriously restricted in range.

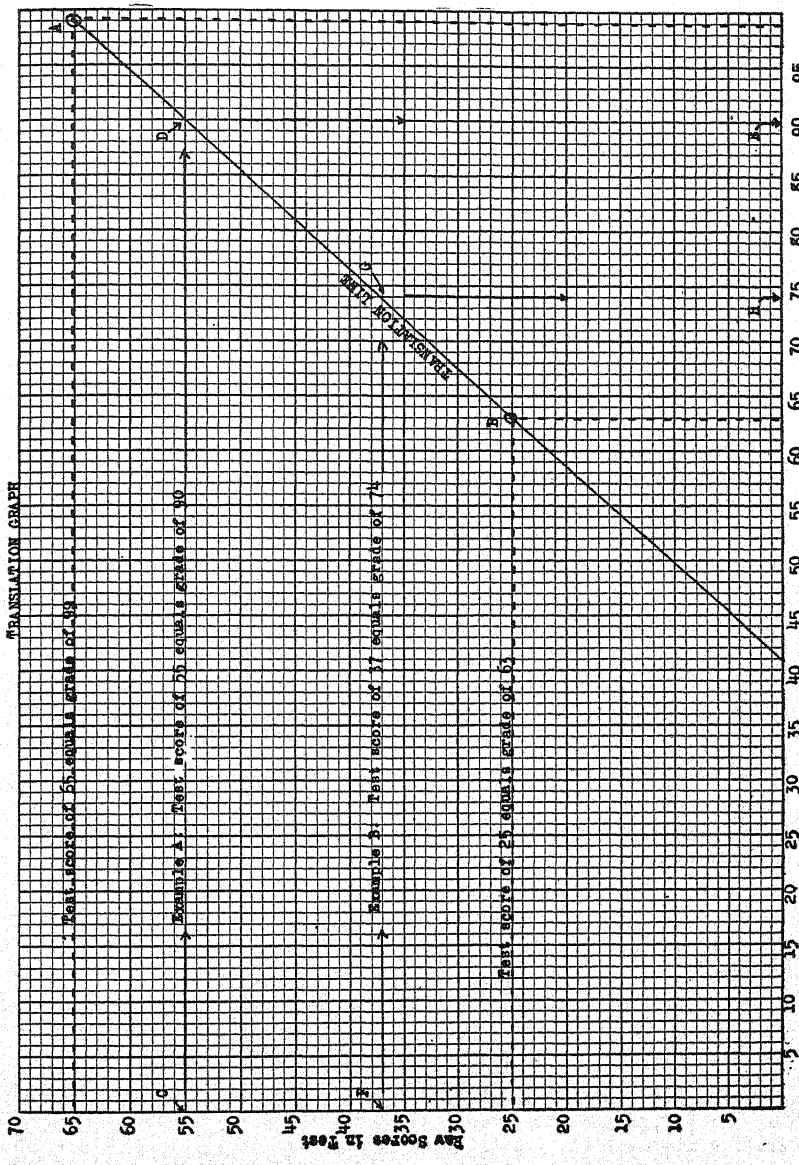


Figure 1-E-2. A translation graph for converting raw scores to Navy Standard Grades.

APPENDIX E-3

CONSTRUCTION AND USE OF ABACS

Abacs are a type of computing diagram designed to decrease the labor involved in statistical analysis. These devices have been found to be useful in expediting some of the statistical work involved in the Navy research program.

There are four basic diagrams in abac methods. In each case one begins with a horizontal and a vertical scale and the appropriate horizontally and vertically lined grid for these scales. Then, upon this grid are superimposed diagonal lines or radial lines.

I. The graph for subtraction is composed of diagonal lines running from lower left to upper right. If one inspects a diagonal connecting the coordinate points (0, 0) and (100, 100), it will be seen that for any point on this diagonal, the value $x - y = 0$. Similarly a diagonal can be drawn such that $x - y = 1$ for any point on the diagonal, etc.

After labelling these diagonals with their $x - y$ values, one can enter the abac with x and y, determine the diagonal on which the coordinate point (x, y) is located, and read the value $x - y$ off the diagonal. An example of a subtraction diagram can be seen in Chart A of abac 4 (estimation of standard error of the biserial correlation coefficient). For further explanation, see text accompanying this abac.

II. The graph for addition is composed of diagonal lines running from lower right to upper left. If one inspects a diagonal connecting the coordinate points (100, 0) and (0, 100), it will be seen that for any point on the diagonal $x + y = 100$. The methods and advantages of the summation graph are the same as those for the subtraction graph.

III. The graph for division is composed of radial lines fanning out from the coordinate point (0, 0). If one wishes to compute the value of x for the equation $x = y/N$, then N values would be indicated by the radial lines. Entering the graph on the y scale one follows the particular horizontal line across to the intersection with the appropriate radial line (N) and from that point follows the vertical line to the x scale and reads off the indicated value. An example of a division diagram can be seen in Chart B of abac 4.

IV. The graph in III could also be used for multiplication, i.e. if $x = y/N$, then $y = xN$. By entering the graph on the x scale, running up to the radial line (N) and across to the y scale, one has multiplied x by N to get the value of y. An example of a multiplication diagram can be seen in Chart C of abac 7 (test of significance of difference between two proportions).

There are several devices which may be utilized to make the basic diagrams of abac methods more useful.

It is frequently possible to set the scale up in terms of a *function* of a parameter and then *label* the scale in terms of *parameter values*. This

eliminates the necessity of computing the value of the function of the parameter for each computation that is made after the original scale has been established. A good example of this is to be found in the formula for the standard error of biserial r.

$$\sigma_{r_{bis}} = \frac{\frac{\sqrt{pq}}{z} - r^2_{bis}}{\sqrt{N}}$$

The $\frac{\sqrt{pq}}{z}$ is a function of p. Therefore it is possible in an abac to construct a scale on the basis of the function $\frac{\sqrt{pq}}{z}$, but to label the scale with the appropriate p-values. Then when using the abac one enters with the p-value and the scaling of the abac automatically takes care of the calculation of the function $\frac{\sqrt{pq}}{z}$.

In the previous discussion we have mentioned only the use of each of the four basic diagrams as a discrete unit. It is frequently necessary to combine in sequence two or more of the basic charts in order to secure a satisfactory abac. Examples of this are to be seen in abacs 3, 4, and 7.

When more than one chart is involved in an abac, it is desirable to use a "continuous line" method to speed up graphic computing. This method consists of placing one chart beside another in an abac so that the answer scale in the first chart also serves as the scale by which one enters the second chart. Under these conditions one does not read the value off the answer scale in the first chart, but instead continues to follow the original line from the first chart into the second chart. An example of this can be seen in the transition from Chart A to Chart B in abac 4. It will be seen that no scale is indicated between the two charts, but instead one simply follows a "continuous line" from one chart to the other. If the two charts were separate, it would be necessary to read off a scale value when coming out of Chart A and then enter Chart B with this scale value. Thus, the "continuous line" method avoids time consuming reference to extra scales and eliminates the possibility of error in transferring from one scale to another.

Because of space limitations it is frequently desirable to expand or contract a scale when going from one chart to another. This is easily done by running an appropriate diagonal across the first of the two charts. See example in Chart A of abac 4.

In order to have a scale cover a sufficiently wide range for general use it is sometimes necessary to compress one end to an undesirable extent. If a high degree of accuracy in the computation of values for that end of the scale is necessary, a second scale can be set up which covers the same distance on the chart as the original scale, but makes use of only that end of the scale which had been undesirably compressed. This technique can be carried out to the number of scales that the person constructing the abac chart considers desirable and efficient.

For the construction of abacs, the use of a "standard cross-section engraving 300 mm." paper approximately 18" x 22" is recommended for generally satisfactory results.

In addition to the four basic diagrams of abac methods, other special abac diagrams may be found useful. Two special diagrams have been used in the abacs presented here.

Abac 6 is an example of the use of one such special diagram. In this abac the diagram is based on the fact that for a point located on the perimeter of a circle the sum of the squares of the coordinates for this point is equal to the square of the radius of the circle (i.e., $r_i^2 = x_{ij}^2 + y_{ij}^2$, where "i" refers to the particular circle and "j" refers to a designated co-ordinate point on the perimeter of this circle).

Another special diagram has been used to compute the value of a weighted average, according to the formula $\bar{P} = \frac{p_1 N_1 + p_2 N_2}{N_1 + N_2}$. This equation can be stated in another form as follows: $\frac{\bar{P} - p_2}{N_1} = \frac{p_1 - \bar{P}}{N_2}$. The ratios of the corresponding sides of similar triangles are equal; therefore the preceding equation can be expressed graphically by similar triangles. Let $(\bar{P} - p_2)$ and N_1 be sides of one triangle, and let $(p_1 - \bar{P})$ and N_2 be corresponding sides of the other triangle, with the $\bar{P}$ scale being the common base line between the similar triangles (see Chart A in abac 7). Then one can read off the value of the weighted average at the point of contact of the two triangles on the $\bar{P}$ scale.

In general, these abacs have been found to have the following advantages over regular computing methods. Using ordinary office personnel, the abacs are faster, less subject to large errors in operation, and require less skill. This results in a large decrease in time required to train a person to use them. In a test comparison of abac methods with machine-computing methods, comparing a highly skilled machine computer on a process with which she was thoroughly familiar with a person who had only two hours' practice in using the abac, the two methods were found to be about equally fast with a small difference in favor of the abac method. This seems to lead to the conclusion that where an experienced machine-computing staff is already operating on routine procedures, there would be no advantage in setting up abacs for these routine procedures. However, where such highly skilled machine computers are not available, or where less routine computations must be carried out for a considerable number of values, it is suggested that the use of abacs should facilitate considerably the work of statistical analysis.

Abac methods are also advantageous where it is desired that students or research personnel be aware of the statistical functions dealt with and the limitations of their procedures. With the graphic method it is easy to notice and demonstrate the effect of changes in the value of a parameter on the answer that will be obtained. It makes one more aware of the degree of accuracy which is desirable in various situations. Abac methods

should be excellent for facilitating the proper orientation of the student toward the statistical tools which are at his disposal.

On the following pages are presented schematic drawings of specific abacs and a discussion of each. Due to space limitations, these sketches do not show the detail and precision which make abacs accurate and efficient to use. Also, in some cases only part of the scale for a variable is shown, while the actual abac presents the entire scale. If more complete information concerning the construction of a specific abac is desired, it is suggested that inquiries be made of the Research Department, College Entrance Examination Board, Princeton, New Jersey.

Abac 1

ABAC TO DETERMINE SIGNIFICANCE OF DIFFERENCE BETWEEN TWO INDEPENDENT PRODUCT-MOMENT CORRELATION COEFFICIENTS (N ASSUMED LARGE)

This abac for testing the significance of difference between two independent product-moment correlation coefficients (N assumed large), is based upon the formula:

$$\text{I. Difference ratio} = \frac{r_1 - r_2}{\sigma_{r_1 - r_2}},$$

$$\text{where } \sigma_{r_1 - r_2} = \sqrt{\frac{1 - \left(\frac{r_1 + r_2}{2}\right)^2}{N/2}}.$$

For convenience in constructing and using the abac, formula I was transformed to the following:

$$\text{II.} \quad (r_1 - r_2)^2 = \frac{(\text{difference ratio})^2}{2N}\,[4 - (r_1 + r_2)^2]$$

By formula II, we determine how large the difference between correlation coefficients ($r_1 - r_2$) must be to be significant at the desired level (in this case the 1% level). One enters the formula with the desired difference ratio (at the 1% level of significance, difference ratio = 2.57582), the sum of the correlation coefficients ($r_1 + r_2$), the number of cases (N), and computes the value of $(r_1 - r_2)^2$. The abac was constructed to avoid the computing labor which is involved in determining the significance of difference between correlation coefficients. Data required to use the abac are r_1, r_2, and N.

To illustrate the use of the abac the following example is presented:

Given the experimentally obtained values $r_1 = .75$, $r_2 = .50$, $N_1 = 200$, and $N_2 = 200$, then $r_1 - r_2 = .25$, $r_1 + r_2 = 1.25$, and $N = 200$. One enters the $r_1 + r_2$ scale at the value 1.25 and follows the vertical coordinate line to the point of intersection with the radial line $N = 200$. Then follow the horizontal coordinate line from this point to the $r_1 - r_2$ scale and read the scale value. For the example given, this value is slightly less than .20. Since the experimentally obtained $r_1 - r_2$ value (.25) is larger than the indicated scale value (.20), the experimentally obtained value is significant at the 1% level. If the experimentally obtained difference had been smaller than the indicated scale value, the

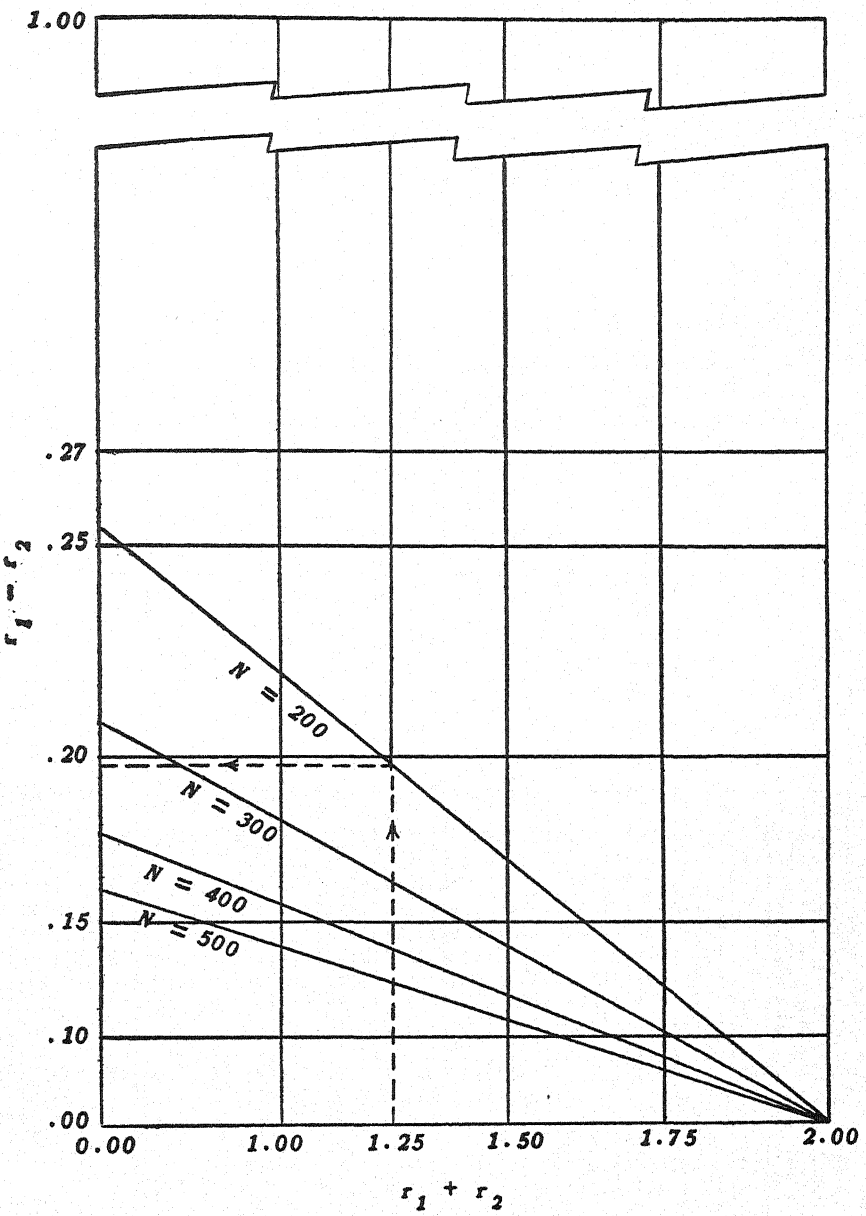
experimentally obtained difference would not be significant at the 1% level. This is assuming that the abac has been constructed to test the significance of differences at the 1% level, as was true in the present case. By using sets of differently colored radial lines, several levels of significance can be indicated on the same abac.

In the example presented above, we have dealt with the special case where the number of persons in each group is the same ($N_1 = N_2$) and $N = N_1 = N_2$. Actually, N is a composite value and for the more general problem where the number of persons is different for the two groups, this composite N should be calculated by the following formula: [1]

$$N = \frac{2N_1N_2}{N_1 + N_2}$$

Since N is assumed large, the composite N is computed as the number of cases rather than the number of degrees of freedom. For computation by abac methods refer to Chart B in abac 7. In the present use of Chart B, the value computed is $\frac{1}{N_1} + \frac{1}{N_2}$ which equals $\frac{2}{N}$. However, by appropriate scaling of line "g" in Chart B, one could read off the value of the composite N from this line.

[1] The calculation of the composite N does not change the manner in which abac 1 is used.



Abac 1. Abac for significance of difference of independent correlations, assuming N to be large.

Abac 2

ABAC FOR TESTING THE SIGNIFICANCE OF DIFFERENCE BETWEEN TWO INDEPENDENT PRODUCT-MOMENT CORRELATION COEFFICIENTS ACCORDING TO FISHER'S Z-TECHNIQUE

Abac 2 is based upon the formula:

$$\text{Diff. ratio} = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Let N_1 represent number of cases in first group,
and N_2 represent number of cases in second group,
then $n_1 = N_1 - 3$, and $n_2 = N_2 - 3$.

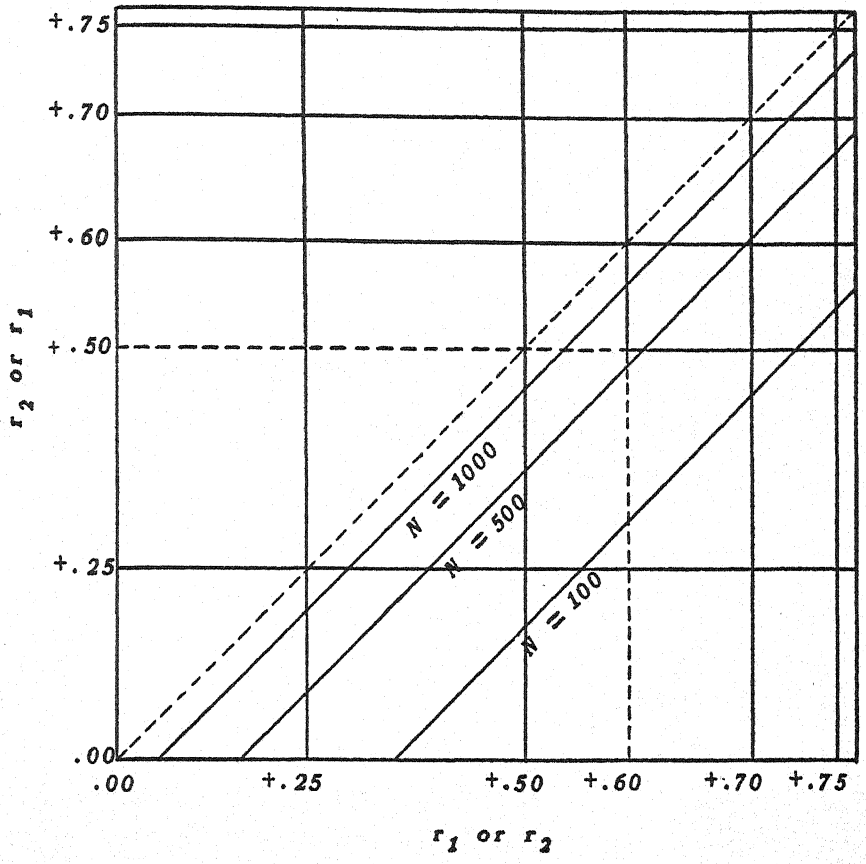
Taking the case where $N = N_1 = N_2$, the formula was transformed to the following working formula:

$$\sqrt{\frac{2}{N-3}} = \frac{z_{r_1} - z_{r_2}}{\text{Diff. ratio}}$$

The abac is constructed to solve for the number of cases which is required for an obtained difference to be significant at a specified level (in the accompanying schematic abac, the 1% level of significance). If the number of cases indicated as necessary by the abac is smaller than the number of cases used in obtaining the experimental difference, then the difference is considered to be significant.

For example, if one had obtained experimentally the following values $r_1 = .60$ and $r_2 = .50$, using 1000 cases in each group ($N_1 = 1000$ and $N_2 = 1000$); then one would enter the abac with the larger r-value ($r_1 = .60$) on the horizontal scale, with the smaller r-value on the perpendicular scale, and determine the coordinate point representing these two values. It can be seen that this point lies between the lines for $N = 500$ and $N = 1000$. This indicates that the number of cases used would have to be slightly greater than 500 for the difference to be significant. Actually, the number of cases used was considerably larger (i.e., $N = 1000$), therefore the difference is definitely considered to be significant. An easy working criterion, in using this abac, is to state that if the determined coordinate point lies below and to the right of the number of cases used (composite N), the difference is considered to be significant at the 1% level. Similar sets of diagonal lines can be drawn for other levels of significance.

If $N_1 \neq N_2$, the value for composite N used in the formula should be computed by a procedure similar to that for composite N in abac 1. Note that for small samples, number of degrees of freedom should be used rather than number of cases.



Abac 2. Abac for testing the significance of difference between two independent product-moment correlation coefficients, according to Fisher's z-technique.

Abac 3

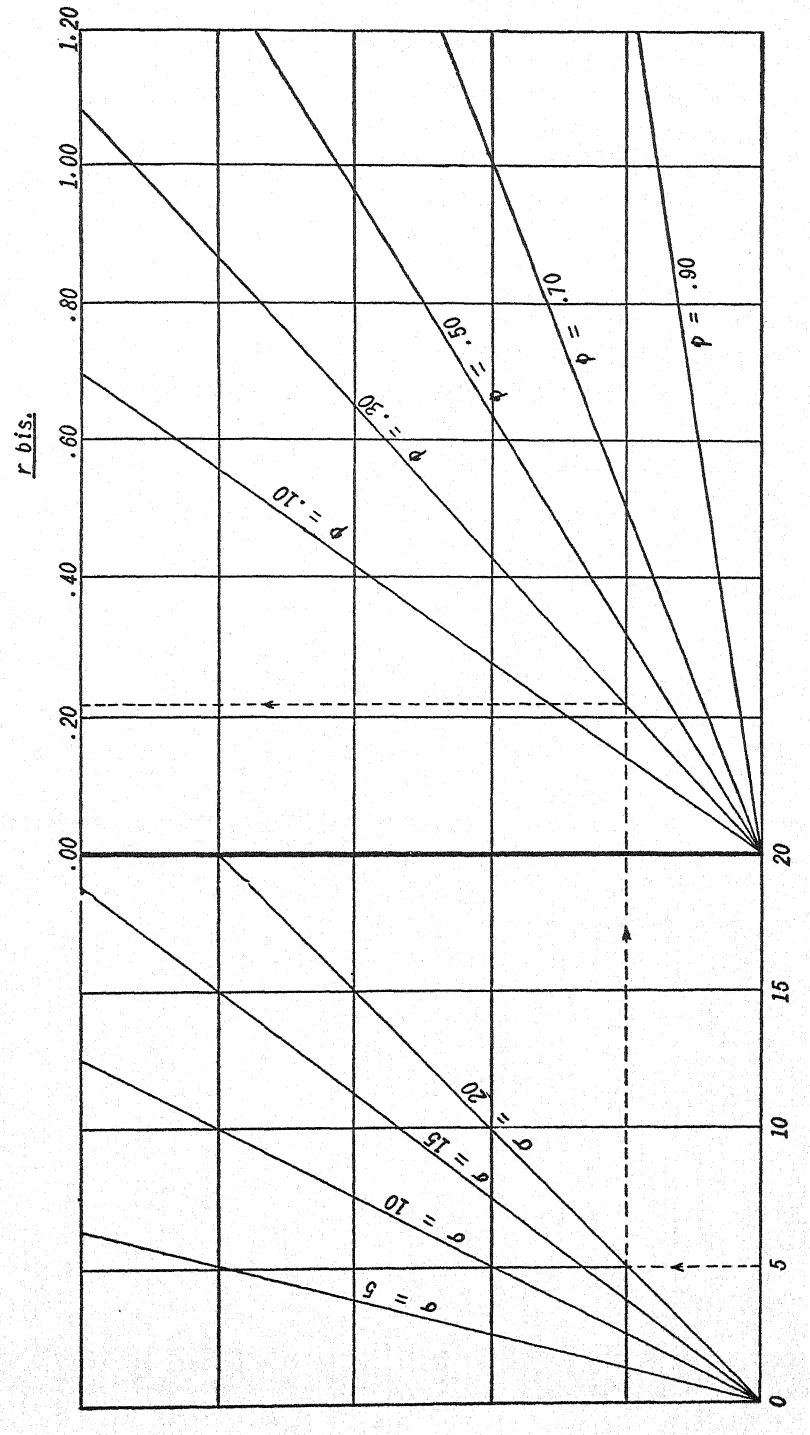
ABAC FOR CALCULATION OF THE BISERIAL CORRELATION COEFFICIENT

This abac is based upon the following formula for the calculation of the biserial correlation coefficient:

$$r_{bis} = \frac{M_+ - M_t}{\sigma_t}\left(\frac{p}{z}\right).$$

Data required to use the abac are: the mean subtest score of the group getting the item correct (M_+); the mean subtest score of the entire group taking subtest (M_t); the standard deviation of the distribution of subtest scores for entire group taking the subtest (σ_t); and the proportion of persons attempting the item who get the correct answer (p).

Given the experimentally obtained values $M_+ - M_t = 5$, $\sigma_t = 20$, and $p = .30$, enter Chart A on the $M_+ - M_t$ scale at 5, run up the vertical line to the radial line for $\sigma_t = 20$, across to the radial line $p = .30$ (in Chart B) and up to the r_{bis} scale, which gives an answer of approximately .21. On a properly constructed abac, the r_{bis} value could be accurately computed to three places.



Abac 3. Abac for the calculation of the biserial correlation coefficient.

Abac 4

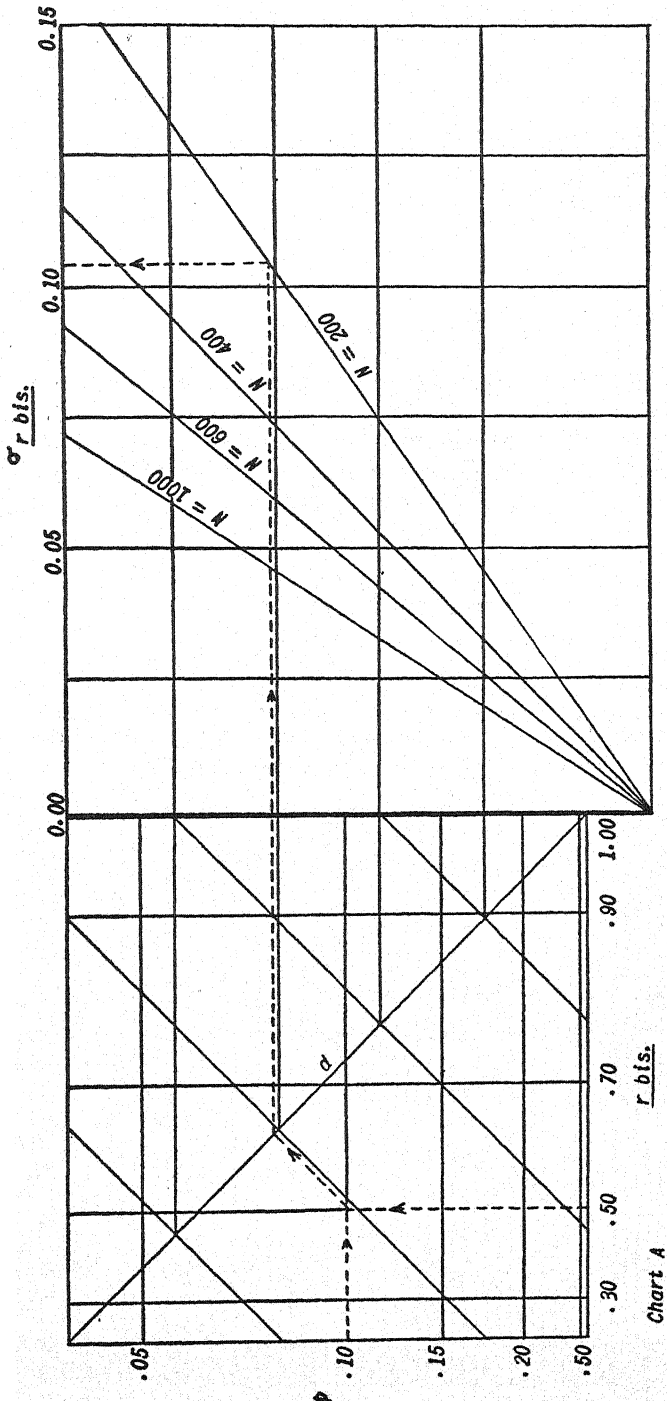
ABAC FOR ESTIMATING STANDARD ERROR OF THE BISERIAL CORRELATION COEFFICIENT

The formula upon which this abac is based is

$$\sigma_{r_{bis}} = \frac{\frac{\sqrt{pq}}{z} - r^2_{bis}}{\sqrt{N}}$$

Data required to use the abac are the difficulty of the item (p), biserial correlation coefficient (r_{bis}), and the number of cases (N). In the abac, Chart A is the subtraction, $\frac{\sqrt{pq}}{z} - r^2_{bis}$. Chart B is the division of this quantity by the $\sqrt{N}$.

Given the experimentally obtained values of $p = .10$, $r_{bis} = .50$, and $N = 200$, one determines in Chart A the coordinate point for $p = .10$ and $r_{bis} = .50$. Then follow the diagonal line to line d, from there follow the horizontal line to the intersection with the radial line, $N = 200$. From the intersection follow the vertical line to the $\sigma_{r_{bis}}$ scale and read off the value. In this case, the value of the standard error of r_{bis} is approximately .103 as determined on our original abac. Machine computations, using eight place figures for $\sqrt{pq}$ and z, gave a value of .1032. It can be seen that properly constructed abacs provide the necessary accuracy. It should be noted that the slanting line d has been placed in the abac to reduce the size of the scale with which one enters Chart B. See general discussion on construction of abacs.



Abac. 4. Abac for estimating standard error of the biserial correlation coefficient.

Abac 5

ABAC FOR ESTIMATING THE POPULATION CORRELATION COEFFICIENT FROM THE COEFFICIENT OBTAINED FROM A CURTAILED SAMPLE

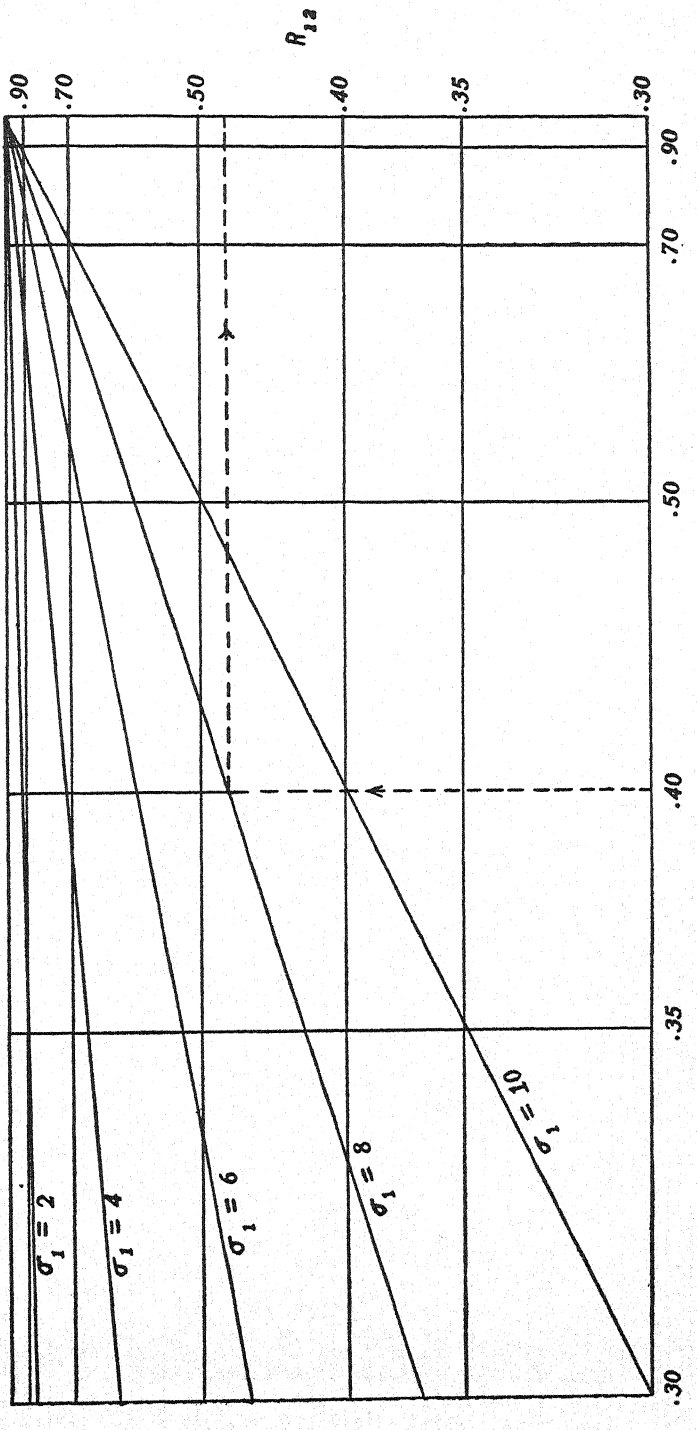
The abac for estimating the correlation coefficient (R) in a standard Navy population ($\Sigma = 10$) from observed r in a curtailed sample is based upon the formula:

$$R_{12} = \frac{r_{12}\frac{\Sigma_1}{\sigma_1}}{\sqrt{1 - r^2_{12} + r^2_{12}\left(\frac{\Sigma_1}{\sigma_1}\right)^2}}$$

Data required to use the abac are the correlation coefficient (r_{12}) and the standard deviation (σ_1) for the curtailed sample. The value of the standard deviation of the standard Navy population is always the same ($\Sigma_1 = 10$). For convenience in using the abac, the preceding formula was transformed to the following working formula:

$$\frac{1}{R^2_{12}} - 1 = \frac{\frac{1}{r^2_{12}} - 1}{\left(\frac{\Sigma_1}{\sigma_1}\right)^2}$$

Given the experimentally obtained values $r_{12} = .40$ and $\sigma_1 = 8$; enter the r_{12} scale at the value .40; run up to the radial line, $\sigma = 8$; across to the R_{12} scale and read off the scale value. In this case the value of the estimated correlation (R_{12}) is approximately .48.



Abac 5. Abac for estimating the correlation coefficient (R) in a standard Navy population from observed r in a curtailed sample.

Abac 6

ABAC FOR ESTIMATING THE STANDARD ERROR OF A SUM OR DIFFERENCE

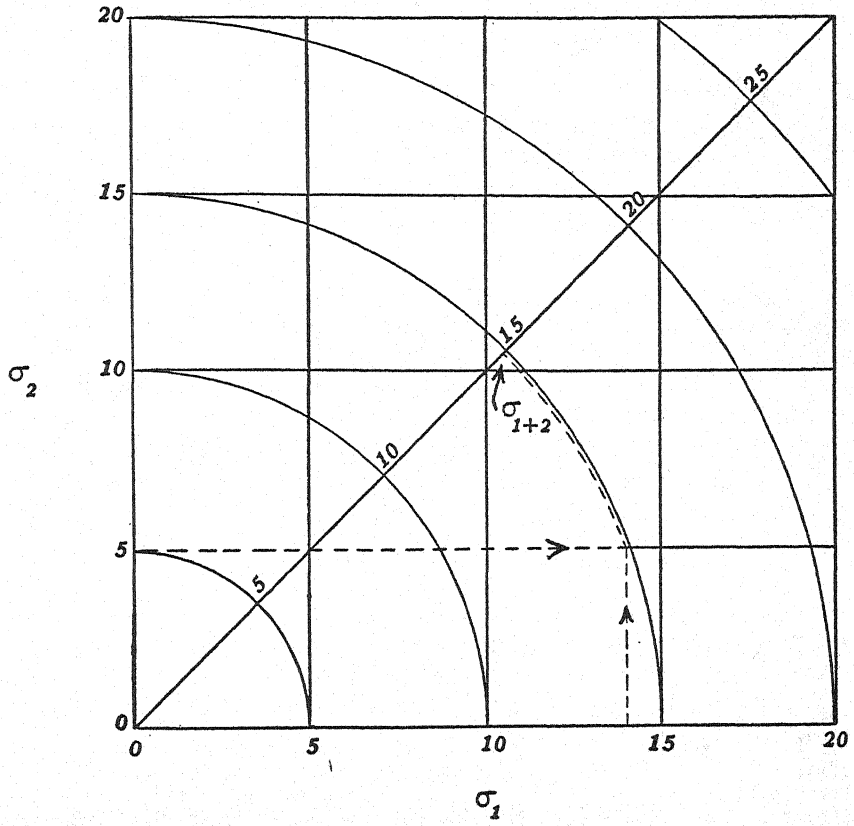
This abac is based upon the formula:

$$\sigma_{1+2} = \sigma_{1-2} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

Data required for using this abac are the standard deviations of the test score distributions (σ_1 and σ_2).

Given the experimentally obtained values $\sigma_1 = 14$ and $\sigma_2 = 5$, determine the coordinate point on the abac which corresponds to these values. From the coordinate point follow the circle to the diagonal and read off the scale value. In this case the answer is approximately 15. The use of good cross section paper in constructing the abac, makes it possible to compute this function graphically for values from 0.000 to 1.000, or any similar range of values.

It should be noted that the use of this formula assumes a zero correlation between variables 1 and 2.



Abac 6. Abac for computing the standard error of a sum or difference, assuming that the correlation between the variables is low.

Abac 7

ABAC FOR TESTING SIGNIFICANCE OF DIFFERENCE BETWEEN TWO PROPORTIONS

The abac for testing the significance of a difference between two proportions is based upon the formula:

$$\text{I.}^{1} \qquad \chi^2 = \frac{(p_1 - p_2)^2/PQ}{\frac{1}{N_1} + \frac{1}{N_2}}$$

where average difficulty value, $P = \frac{p_1 N_1 + p_2 N_2}{N_1 + N_2}$ and $Q = 1 - P$.

Data required to use the abac are, for each of the two items, the number (N) attempting it and the proportion (p) of those attempting it who answer correctly.

For convenience in using the abac, the formula was transformed to the following:

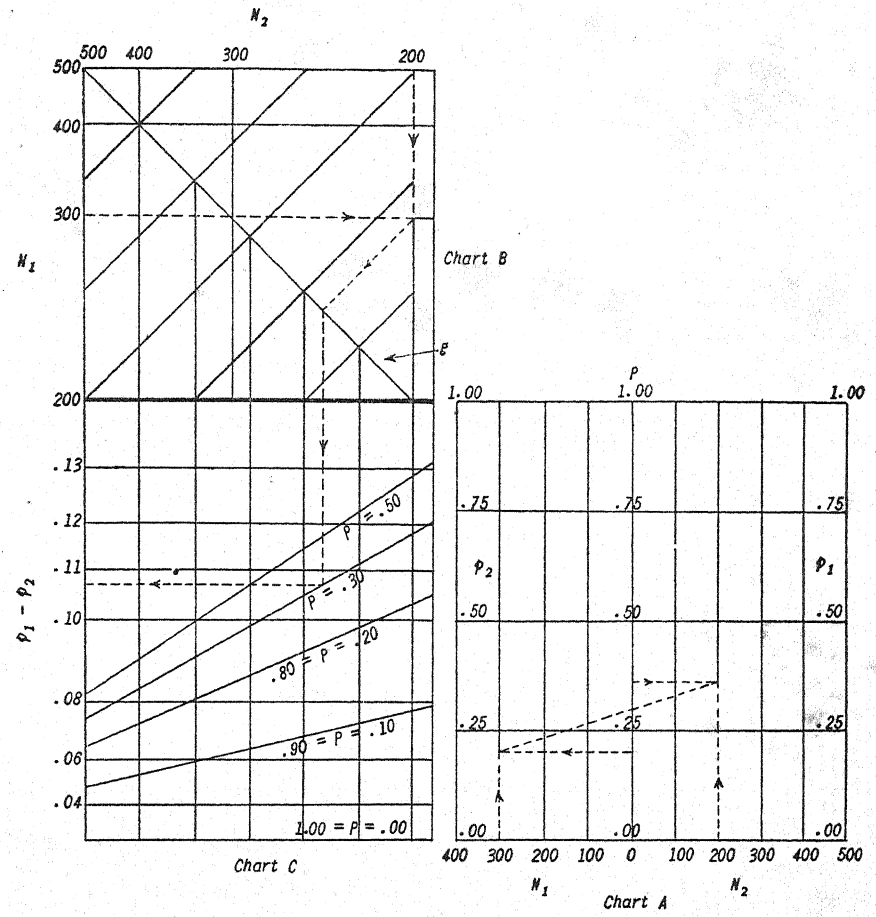
$$\text{II.} \qquad (p_1 - p_2)^2 = (\chi^2)\,(PQ)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)$$

By formula II we determine how large the difference between p-values must be to be significant at the desired level (in this case the 1% level). The abac has been constructed to perform this calculation.

To illustrate the use of the abac, the following example is presented:

Note that there are three charts, A, B, and C. Given the experimentally obtained values $p_1 = .37$, $p_2 = .20$, $N_1 = 300$, and $N_2 = 200$, one enters Chart A on the abac and determines the coordinate points (p_2, N_1) and (p_1, N_2). Connect these two points by a straight line and read off the P value at the point where the connecting line crosses the P scale. In this case, the value is .30. Then mark the radial line in Chart C for $P = .30$. In actually using abacs, it is desirable to have pivoted rulers which can be swung around to the appropriate radial line. Now enter Chart B and find the coordinate point (N_1, N_2). From this point move diagonally to line "g" and then vertically into Chart C to the previously marked radial line $P = .30$. From there, run across horizontally to the ($p_1 - p_2$) scale and read off the value, which is approximately .108. Compare the experimentally obtained p difference (.37 − .20 = .170) with the theoretical difference (.108) just read off the scale. Since the experimentally obtained difference is larger than the theoretical difference which is indicated as being significant at the 1% level, the experimental difference is significant. Conversely, if the experimental difference had been smaller than the theoretical difference read off the scale, the experimental difference would *not* be significant at the 1% level. By drawing other sets of radial lines in Chart C one could test the difference at any desired level of significance (i.e., the position of the sets of radial lines is determined as a function of the value of χ^2 which is selected).

[1] For one degree of freedom, $\chi^2 = 6.635$ at the 1% level of significance.



Abac 7. Abac for testing significance of differences between two proportions.

APPENDIX E-4

TABLE FOR ESTIMATING POPULATION CORRELATION COEFFICIENT FROM CORRELATION COEFFICIENT OBTAINED ON A RESTRICTED SAMPLE

Kelley's formula for estimating the correlation coefficient in an unrestricted population from the correlation coefficient in a restricted sample (from which abac 5, Appendix E-3, has been constructed) may also be presented in the form of a table. The accompanying table was actually made from readings on such an abac and is therefore subject to an error of $\pm .005$. The column headings are values of $\frac{\sigma_1}{\Sigma_1}$, the ratio of the standard deviation in the restricted sample to the standard deviation in the population. The row headings are the correlation coefficients obtained in the restricted sample. The estimated correlation coefficient in the population is found in the body of the table at the intersection of the appropriate row and column. To save space, all decimal points have been omitted.

Given the experimentally obtained values of $r = .40$ in the restricted sample; $\sigma = 8.0$, the standard deviation in the sample; $\Sigma = 10.0$, the standard deviation in the population; then the table is entered at the column headed 80 and the row headed 40. At the intersection of that column and that row is 48. The desired estimated correlation coefficient is .48.

Discussion of the limitations in the use of this estimation is found in Chapter XII.

Ratio of Sigma of Sample to Sigma of Population

Observed Coefficient of Correlation

	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
60	83	83	82	82	81	81	80	80	79	78	78	77	77	77	76	76	75	75	74	74	73	73	72	72	71	71	70	70	69	69	68	68	68	67	67	66	66	65	65	65	64	64	63	63	62	62	62	61	61	60
59	82	82	81	81	80	80	79	79	78	78	77	77	76	76	75	75	74	74	73	73	72	72	71	71	70	70	69	69	68	68	68	67	67	66	66	65	65	64	64	64	63	63	62	62	62	61	61	60	60	59
58	82	81	81	80	80	79	79	78	77	77	76	76	75	75	74	74	73	73	72	72	71	71	70	70	69	69	68	68	67	67	67	66	66	65	65	64	64	63	63	63	62	62	61	61	60	60	60	59	59	58
57	81	80	80	80	79	78	78	77	77	76	76	75	75	74	74	73	73	72	71	71	70	70	69	69	68	68	67	67	66	66	65	65	64	64	63	63	62	62	62	61	61	60	60	59	59	59	58	58	57	57
56	80	80	79	79	78	78	77	76	76	75	75	74	74	73	73	72	72	71	71	70	70	69	69	68	68	67	67	66	66	65	65	64	64	63	63	62	62	61	61	61	60	60	59	59	58	58	58	57	57	56
55	80	79	78	78	77	77	76	76	75	75	74	73	73	72	72	71	71	70	70	69	69	68	68	67	67	66	66	65	65	64	64	63	63	62	62	61	61	60	60	59	59	59	58	58	57	57	57	56	56	55
54	79	78	78	77	76	76	75	75	74	74	73	73	72	71	71	70	70	69	69	68	68	67	67	66	66	65	65	64	64	63	63	62	62	61	61	60	60	59	59	58	58	58	57	57	56	56	56	55	55	54
53	78	77	77	76	76	75	75	74	73	73	72	72	71	70	70	69	69	68	68	67	67	66	66	65	65	64	64	63	63	62	62	61	61	60	60	59	59	58	58	57	57	57	56	56	55	55	55	54	54	53
52	77	77	76	75	75	74	74	73	72	72	71	71	70	70	69	68	68	67	67	66	66	65	65	64	64	63	63	62	62	61	61	60	60	59	59	58	58	57	57	56	56	56	55	55	54	54	54	53	53	52
51	76	76	75	75	74	73	73	72	71	71	70	70	69	69	68	67	67	66	66	65	65	64	64	63	63	62	62	61	61	60	60	59	59	58	58	57	57	56	56	55	55	55	54	54	53	53	53	52	52	51
50	76	75	74	74	73	72	72	71	71	70	69	69	68	68	67	67	66	65	65	64	64	63	63	62	62	61	61	60	60	59	59	58	58	57	57	56	56	55	55	54	54	54	53	53	52	52	52	51	51	50
49	75	74	73	73	72	71	71	70	70	69	68	68	67	67	66	66	65	64	64	63	63	62	62	61	61	60	60	59	59	58	58	57	57	56	56	55	55	54	54	53	53	53	52	52	51	51	51	50	50	49
48	74	73	73	72	71	71	70	69	69	68	67	67	66	66	65	65	64	63	63	62	62	61	61	60	59	59	58	58	57	57	56	56	55	55	54	54	53	53	52	52	52	52	51	51	50	50	50	49	49	48
47	73	72	72	71	70	70	69	68	68	67	66	66	65	65	64	64	63	62	62	61	61	60	60	59	58	58	57	57	56	56	55	55	54	54	53	53	52	52	51	51	51	51	50	50	49	49	49	48	48	47
46	72	71	71	70	69	69	68	67	67	66	65	65	64	64	63	62	62	61	61	60	60	59	58	58	57	57	56	56	55	55	54	54	53	53	52	52	52	51	51	50	50	49	49	49	48	48	47	47	47	46
45	71	70	70	69	68	68	67	66	66	65	64	64	63	63	62	61	61	60	60	59	58	58	57	57	56	56	55	55	54	54	53	53	52	52	51	51	51	50	50	49	49	48	48	48	47	47	46	46	46	45
44	70	69	69	68	67	67	66	65	65	64	63	63	62	61	61	60	60	59	58	58	57	57	56	56	55	55	54	54	53	53	52	52	51	51	50	50	50	49	49	48	48	47	47	47	46	46	46	45	45	44
43	69	68	68	67	66	66	65	65	64	63	63	62	61	60	60	59	59	58	57	57	56	56	55	55	54	54	53	53	52	52	51	51	50	50	49	49	49	48	48	48	47	47	46	46	46	45	45	44	44	43
42	68	67	67	66	66	65	64	64	63	62	62	61	60	60	59	59	58	57	57	56	56	55	55	54	54	53	53	52	52	51	51	50	50	49	49	48	48	47	47	47	46	46	45	45	45	44	44	43	43	42
41	67	66	65	65	64	63	63	62	61	61	60	59	59	58	58	57	56	56	55	55	54	53	53	52	52	51	51	50	50	49	49	48	48	47	47	47	46	46	45	45	45	44	44	44	43	43	42	42	42	41
40	66	65	64	63	63	62	62	61	60	59	59	58	58	57	56	56	55	55	54	53	53	52	52	51	51	50	50	49	49	48	48	47	47	47	46	46	45	45	44	44	44	43	43	42	42	42	41	41	41	40
39	65	64	63	62	62	61	60	60	59	58	58	57	56	56	55	55	54	53	53	52	52	51	51	50	50	49	49	48	48	47	47	46	46	45	45	45	44	44	43	43	43	42	42	41	41	41	40	40	40	39
38	63	63	62	61	61	60	59	58	58	57	56	56	55	55	54	53	53	52	52	51	51	50	50	49	49	48	48	47	47	46	46	45	45	44	44	43	43	42	42	42	41	41	40	40	40	39	39	39	38	38
37	62	62	61	60	59	59	58	57	57	56	55	55	54	53	53	52	52	51	51	50	49	49	48	48	47	47	46	46	45	45	44	44	43	43	42	42	41	41	41	40	40	40	39	39	39	38	38	38	37	37
36	61	60	60	59	58	57	57	56	55	55	54	53	53	52	52	51	50	50	49	49	48	48	47	47	46	46	45	45	44	44	43	43	42	42	41	41	41	40	40	40	39	39	39	38	38	38	37	37	37	36
35	60	59	58	58	57	56	55	55	54	54	53	52	52	51	50	50	49	49	48	48	47	47	46	46	45	45	44	44	43	43	42	42	41	41	41	40	40	39	39	39	38	38	38	37	37	37	36	36	36	35
34	59	58	57	56	55	55	54	53	53	52	51	51	50	50	49	49	48	47	47	46	46	45	45	44	44	43	43	42	42	41	41	41	40	40	39	39	39	38	38	38	37	37	37	36	36	36	35	35	35	34
33	57	56	56	55	54	54	53	52	51	51	50	50	49	48	48	47	47	46	46	45	45	44	44	43	43	42	42	41	41	40	40	39	39	38	38	38	37	37	36	36	36	35	35	35	34	34	34	34	33	33
32	56	55	54	54	53	52	51	51	50	49	49	48	48	47	47	46	46	45	44	44	43	43	42	42	41	41	40	40	40	39	38	38	38	37	37	36	36	36	35	35	35	34	34	34	33	33	33	33	32	32
31	55	54	53	52	51	51	50	49	49	48	48	47	47	46	45	45	44	44	43	43	42	42	41	41	40	40	39	39	39	38	38	37	37	36	36	36	35	35	35	34	34	34	33	33	33	32	32	32	32	31
30	53	52	52	51	50	50	49	48	48	47	46	46	45	45	44	43	43	42	42	41	41	40	40	39	39	39	38	38	37	37	36	36	36	35	35	35	34	34	34	33	33	33	32	32	32	31	31	31	31	30
29	52	51	50	50	49	48	47	47	46	46	45	44	44	43	43	42	42	41	41	40	40	39	39	38	38	37	37	37	36	36	35	35	35	34	34	34	33	33	33	33	32	32	32	31	31	31	30	30	30	29
28	50	49	49	48	47	47	46	45	45	44	44	43	42	42	41	41	40	40	39	39	38	38	37	37	36	36	35	35	34	34	34	33	33	33	32	32	32	31	31	31	30	30	30	30	29	29	29	29	28	28
27	49	48	47	47	46	45	45	44	43	43	42	42	41	41	40	40	39	39	38	38	37	37	36	36	35	35	34	34	33	33	33	33	32	32	31	31	31	30	30	30	29	29	29	29	28	28	28	28	27	27
26	47	47	46	45	45	44	43	43	42	42	41	40	40	39	39	38	38	37	37	36	36	35	35	35	34	34	33	33	33	32	32	32	31	31	31	30	30	30	29	29	29	28	28	28	28	27	27	27	27	26
25	46	45	44	44	43	43	42	41	41	40	40	39	39	38	38	37	36	36	36	35	35	35	34	34	33	33	32	32	32	31	31	30	30	30	29	29	29	28	28	28	28	27	27	27	27	26	26	26	25	25
24	44	44	43	42	42	41	41	40	39	39	38	38	37	37	36	36	35	35	34	34	33	33	33	32	32	32	31	31	30	30	30	29	29	28	28	28	28	27	27	27	27	26	26	26	25	25	25	25	24	24
23	43	42	41	41	40	39	39	38	38	37	37	36	36	35	35	34	34	33	33	32	32	32	31	31	30	30	30	29	29	29	28	28	28	27	27	27	27	26	26	26	25	25	25	25	24	24	24	24	23	23
22	41	40	40	39	38	38	37	37	36	36	35	35	34	34	33	33	32	32	31	31	31	30	30	30	29	29	28	28	28	27	27	27	26	26	26	25	25	25	25	24	24	24	24	24	23	23	23	23	22	22
21	39	39	38	38	37	36	36	35	35	34	34	33	33	32	32	31	31	31	30	30	29	29	28	28	28	27	27	27	26	26	26	25	25	25	24	24	24	24	23	23	23	23	23	22	22	22	22	22	21	21
20	38	37	37	36	35	35	34	34	33	33	32	32	31	31	30	30	30	29	29	28	28	28	27	27	27	26	26	26	25	25	25	24	24	24	24	23	23	23	23	22	22	22	22	21	21	21	21	21	20	20
	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99

INDEX

abacs, construction and use, 488-505
achievement examinations (see also under tests, achievement), list, 461-464
achievement testing program, development, 287-290; results, 303-313; scope, 288-290
adjustment (see personal adjustment)
advancement in rating, procedures, 45, 46, 346-347; qualifications, 344-347
advancement in rating examinations (see also tests), list, 460-461
age and test scores, (*Basic Test Battery*), 99, 103
age as a predictive measure (see under predictive measures)
Aid-to-the-Interview Blank, 23
Airplane Matching Test, 124
Applicant Qualification Test (see under tests)
Applied Psychology Panel, 8
aptitude tests (see also tests, classification), needed research, 434 to 438
Arithmetical Computation Test, 77-78
Arithmetical Reasoning Test (see tests, *Arithmetical Reasoning* and *Basic Test Battery*)
Army-Navy College Qualifying Test (see tests, *College Qualifying*)
attitude studies (see information surveys)

Basic Test Battery (see tests)
Billet Preference Record, 135
Billet Qualifications Blank (see tests)
Billet Selection Requirements Manual, 16
biographical information, research on use in classification, 440
Biographical-Preference Inventory, 227, 232
Bureau of Medicine and Surgery, 128
Bureau of Navigation, 7

C-1 Test (see tests, *College Qualifying*)
CIC Aptitude Test (see tests)
CIC Final Achievement Examination (see tests)
civilian education (see predictive measures)
civilian occupation (see predictive measures)
College Entrance Examination Board, contracting agency, 8; staff members, 458; Navy research reports, 468-469
College Qualifying Test (see tests, *College Qualifying*)
combat fatigue, 126
combinations of test scores for prediction of success, 243-252, 261
Committee on Selection and Training of Service Personnel, 8
Constructing and Using Achievement Tests, 293
Cornell Selectee Index, 131
course grades, reliability, 183, 188-189, 197
criteria, Navy, evaluation, 377-379, 392-393
criterion, acceptable characteristics, 358-362; improvement, 293, 296, 306-307, 329-330; limitations, 138-141, 178, 222-223, 268-272, 392-393; meaning of term, 357; of personal adjustment, 138-140, 174; of proficiency—nomination ratings, 363, 386; of success in training, achievement test scores, 202-211, 220, 293, 296, 306-307, 329-330, 378; class rank, 219, 228; course grades: officer, 178, 180, 186-187, 191-193, 202, 206-210, 219-220, 327-329, 376, 377; enlisted, 233, 236, 268-272, 305; pass-fail, 192, 196; problems in development 362-376; supervisors' ratings, 220, 378-379 (see also *Officer-Aptitude Rating*); psychiatric, 138-140
curriculum analysis, 333

Digit Memory Span Test, 124
Duty Recommendation Form, 16, 18, 217

Enlisted Personal Inventory, (see tests)
enlisted personnel, classification, 22-28; procurement, 21-22; recommendation for assignment, 23, 234; records, 28-29; tests, (see also Chapters VI, VIII, XII, XIII, XV, XVII, XVIII, XX), list, 459-464; training, advancement in rating, 45-46; operational, 45; recruit, 42-44; service schools advanced: description, 44-45, 262-264, prediction of success, 262-284; elementary service schools: description 44, prediction of success, 233-261; special programs, 46-49; WAVES, 47
Enlisted Personnel Qualifications Card, 22-23, 236, 252, 267, 384
examinations (see tests)
experience (see predictive measures)
Experience Comparison Index (see tests)
Eye-Hand Coordination Test, 123-124

factor analyses, *Basic Test Battery,* 74-75
failure in training and *General Classification Test Scores,* 257-258

fleet schools (see operational training under enlisted personnel and under officer personnel)
fleet validation (see shipboard performance)
forced-choice format, 130

General Classification Test (see tests)
General Mathematics Test (see tests)
General Physics Test (see tests)
grading systems, improvement, 292-294, 296, 305-306, 327-329
grading system, Navy, 318

identification tests (see tests, achievement)
indoctrination school, description of program, 35; prediction of success, 177, 180-186, 211-212
Information Surveys, amphibious forces, 421-429; off-duty education (educational services), 415-421; research methodology, 429-432; training, 410-415
interest measures, needed research on use in classification, 439-440
interview (see predictive measures), enlisted, 22-23; needed research, 440-441; officer, 14-17; validity, 251-254, 260, 383, 405
interviewing officers, training, 15
inventories (see personal adjustment measures and biographical information)

job requirements, needed research, 434-435

Literacy Test, 78-80
literacy training program, tests, 78-81

Manual of Enlisted Navy Job Classifications, 26, 29
Mechanical Aptitude Test (see tests)
Mechanical Comprehension Test, 124
Mechanical Knowledge Test (see tests)
morale studies (see *Information Surveys*)
motivation of trainees, needed research, 448-450

National Defense Research Committee, 7, 8
Navy standard score, 103
N-4 Test (see tests)
NDRC Project N-106, initiation, 8; research reports, 468-469; staff members, 458
nominating technique, 363, 386
Non-Verbal Classification Test, 80-81
NROTC Selective Examination (see tests)
NROTC training (see officer personnel, training, college)

Office of Scientific Research and Development, 7, 8
Officer Aptitude Rating, 180, 182, 185, 192
Officer Classification Battery, 107-108
Officer Classification Test (see tests)
Officer Personal Inventory (see tests)
officer personnel, classification procedures, 14-19, 216-217; records, 19; procurement, 12-14; tests, (see Chapters VII, VIII, X, XI, XVI), list, 459, 461; training programs; administration, 31, advanced, description, 37-38, 216-219, prediction of success, 223-228, instructors, 315-316, operational, description, 38-39, 218-219, prediction of success, 228-230, primary schools, (see also indoctrination schools, Reserve Midshipmen's Schools, college training program, V-12 training), description, 34-37; Regular Navy, 32-33; training school reports, 217
Officer Qualification Test (see tests)
Officer Qualifications Questionnaire, 19
Officer Qualifications Record Jacket, 19, 217
Officers' Selective Examination (see tests, *Officer Qualification Test*)
operational fatigue, 126
opinion studies (see *Information Surveys*)
Ortho-Rater Tests, 24

performance in billets (see shipboard performance)
performance tests (see tests, achievement)
personal adjustment, 126
personal adjustment measures, criterion for validation, 138-140, 174; hierarchy of scores, 167-169; item format, 132, 170-172; item difficulty, 171-172; procedures in development, 136-138; research design for validation, 140-141, 173-174; uses and misuses, 127-128; use in classification, 439; validation studies, 141-169; validity of item content, 169-170
Personal Check List, (see tests)
personal inventories (see personal adjustment measures and biographical information)
personality evaluation (see personal adjustment measures)
prediction of success (see also enlisted personnel, training; officer personnel, training; shipboard performance), improvement, 184-186, 190, 199-200, 205-209
Predictive measures (see under name of test or rating measure, and Chapters X, XI, XII, XIII, XX), age, 227, 231, 235-236, 254, 260, 279-280, 383, 399;

civilian education, 227, 231, 235-236, 254-256, 260-261, 280, 383, 399-400; civilian occupation, 227-232, 282-283, 383; effort, 404-405; interview, 251-254; limitation, 267-268; months of active duty, 273, 279; naval experience, 383, 395-396; naval training, 280-282, 383, 400-404; Navy pay grade, 279, 383
Pre-Radar Final Achievement Examination, 320-322
Pre-Radar General Aptitude Test (see tests, *Pre-Radar Officer Aptitude Test Battery*)
Pre-Radar Officer Aptitude Test Battery (see tests)
Previous Duty Check List, 135
proficiency tests, research on use in classification, 438
psychiatric criterion, 138-140
psychiatric interview, 127
psychiatric screening tests (see personal adjustment measures)

"Quality Classification" (see interview), description, 235-236; validity, 251-254, 260
questionnaires (see *Information Surveys* and personal adjustment measures)

Radio Code Test—Speed of Response (see tests)
Radio technician, selection, 331-332; training, 332; qualifications, 344-346
Radio Technician Selection Test (see tests)
rating, Navy (see advancement in rating)
rating scales, radio technician program, 340-342
ratings, supervisors (see criterion of success in training)
Reading Achievement Examination, 290
Reading Classification Examination, 290
Reading Test (see tests)
records, enlisted personnel, 28-29; officer personnel, 19
recruit training (see enlisted personnel training)
referral scores, 127
Relative Movement Test, 107, 112-114
reliability (see criterion, acceptable, and under name of test or battery)
research methodology (see under criterion and prediction of success), 136-141, 138-141, 169-174, 429-432, 450-453
research studies, list, 465-469
Reserve Midshipmen's School, description of program, 34
Reserve Midshipmen's School (Deck), prediction of success, 177, 191-201, 212-213; *Standardized Examination,* 319-320, 323-326
Reserve Midshipmen's School (WR), (see under WAVES officer training)
restriction in range, correction, 237-238, 506-507

school grades (see course grades)
Screening tests (see personal adjustment measures)
Selectometer, 24
self-idealizational scale, 133
Sentences in Noise Test, 124-125
Service Record, 28
service schools (see enlisted personnel, training)
sex differences in test scores, 90-91, 98-99
shipboard performance, criterion, 384-393, 405-406; nature, 380-381; predictive measures, 382-384, 393-409
Social Judgments Test, 136
Sonar Pitch Memory Test, 24, 123
special aptitude tests, descriptions and uses, 112-125; needed research, 436-439
Spelling Test (see tests)
stop item, in measures of adjustment, 131
suppressor variable, 133, 173

tactical radar, (see *CIC Aptitude Test* and *CIC Final Achievement Examination*)
Telephone Talker Test, 24
Test and Research Section, history, 611; mission and accomplishment, 3-6; organization, 4-5, 457; research studies, 465-467; staff members, 457
test item file, 351-353
test item format (see descriptions of tests)
test scores, sex differences, 90-91, 98-99
tests:
achievement (see also training); as criterion (see criterion); conversion of scores, 485-487; enlisted personnel, 295-314, 462-464; identification, 298-299, 301-303, 476-481; list, 461-464; officer personnel, 315-330, 461; performance, 297-298, 300-301, 340, 470-475; procedures used in development, 299-303, 316-318, 334-340; radio technician training, 331-340; types, 288, 296-299, 318-320, 334-335; uses (see also criterion, achievement tests) 288, 290-296, 320-330; value for predicting success in training, 337
advancement in rating, cutting scores, 353-354; development, 348-354; list, 460-461; scope, 347-348;
Airplane Matching Test, 124
Applicant Qualification Test, description, 76-77; use, 22

tests—*continued*
aptitude (see below under tests, classification)
Arithmetical Computation Test, 77-78
Arithmetical Reasoning Test (see also test, *Basic Test Battery*), description, 57
Army-Navy College Qualifying Test (see below, *College Qualifying Test*)
Basic Test Battery (including *Arithmetical Reasoning Test, Clerical Aptitude Test, General Classification Test, Mechanical Aptitude Test, Mechanical Knowledge Test, Radio Code Test—Speed of Response, Reading Test, Spelling Test*), comparison of results at stations, 66; cuting scores, 233; development, 54-55; development of norms, 65-66; description, 55-63; factor analyses, 74-75; intercorrelations, 72-74; item analyses, 64; relation of scores to age, 71-72; relation of scores to highest school grade completed, 71-72; reliability, 68-71; revision, 82-83, 435-436; time limit study, 64-65; time trends in scores, 66, 68; uses, 22, 61-63; validity for classification, service schools elementary, 75, 236, 240-252, 258-260, service schools advanced, 267, 275-279; validity for shipboard performance, 383, 397-399
Billet Preference Record, 135
Billet Qualifications Blank, description, 132-134; research studies, 153-162
Biographical-Preference Inventory, 227, 232
C-1 Test (see below, *College Qualifying Test*)
CIC Aptitude Test, description, 112-115; validity, 115-116, 226-227
CIC Final Achievement Examination, 323, 327-329
classification tests, list, 459-461
Clerical Aptitude Test (see also tests, *Basic Test Battery*), description, 60-61
College Qualifying Test, description, 108-109, 201-202; norms, 110; statistical analyses, 110; uses, 84, 111; validity, 203-209, 213-214
Cornell Selectee Index, 131
Digit Memory Span Test, 124
Enlisted Personal Inventory, description, 130-131; research studies, 142-153, 166-167
Experience Comparison Index, description, 134; research study, 161-163
Eye-Hand Coordination Test, 123-124
General Classification Test (see also tests, *Basic Test Battery*), description, 56; and failure in training, 257
General Mathematics Test (see also *Pre-Radar Officer Aptitude Test Battery*), validity, 223-226
General Physics Test (see also *Pre-Radar Officer Aptitude Test Battery*), validity, 223-226; in use prior to 1943 enlisted personnel), 6, 53
Literacy Test, 78-80
Mechanical Aptitude Test (see also tests, *Basic Test Battery*), description, 57-59
Mechanical Comprehension Test, 124
Mechanical Knowledge Test (see also tests, *Basic Test Battery*), description, 59
N-4 Test, description, 191, 202; validity for prediction, 193-200; as a criterion of success in training, 202-211, 212-214
Non-Verbal Classification Test, 80-81
NROTC Selective Examination, 191-200, 213
Officer Classification Battery, 107-108
Officer Classification Test, age and test scores, 103; description, 101-102; development, 100-101; development of norms, 103; revision, 106-107; statistical analyses, 104; uses, 16, 84, 102-103; validity, 104-106, 112, for primary training, 193-201, 213, for advanced training, 112, 223-228, 230, 329-330, for operational training, 228-231, 230-231
Officer Personal Inventory, description, 131; validity, 227, 232
officer personnel (see officer personnel)
Officer Qualification Test, age and test scores, 99; comparison of scores made by different groups, 90-91, 97-99; description, 85-87; development, 85, 482-484; development of norms, 88; sex differences and test scores, 90-91, 98-99; statistical analyses, 88-92, 96; stability of items, 97; use, 12, 84, 87-88; validity, 92-96, 180-186, 186-190, 193-200, 211-212
Personal Check List, description, 134-135; research study, 163-167
Personal Inventory, 129
Personal Inventory, Form X-1(20), 131-132
Pre-Radar Final Achievement Examination, 320-322
Pre-Radar General Aptitude Test (see *Pre-Radar Officer Aptitude Test Battery*)
Pre-Radar Officer Aptitude Test Battery (see also *General Mathematics Test* and *General Physics Test*), 116-118
Previous Duty Check List, 135
Radio Code Test—Speed of Response (see also *Basic Test Battery*), use, 112; description, 121-123

tests—*continued*
Radio Technician Selection Test, 118-121
Reading Achievement Examination, 290
Reading Classification Examination, 290
Reading Test (see tests; *Basic Test Battery*), 56-57
Relative Movement Test, 107, 112, 113, 114
reliability (see name of test or battery)
Reserve Midshipmen's Schools (Deck), Standardized Examination, 319-320, 323-326
selection tests (see tests, classification)
Sentences in Noise Test, 124-125
Social Judgments Test, 136
Sonar Pitch Memory Test, 24, 123
Spelling Test (see tests, *Basic Test Battery*), description, 61
Telephone Talker Test, 24
Winchman and Hatchman Selection Test, 123
training (see also enlisted personnel and officer personnel), evaluation, 309-311, 323-326; failure and *General Classification Test* scores, 257; improvement, 290-292, 296, 303-313, 322, 332-334, 339; needed research, 442-450

validation (see criterion; enlisted personnel, training; officer personnel, training, prediction of success; predictive measures; shipboard performance; and name of individual test or battery)
V-12 training, prediction of success, 177, 201-211, 213-214; description of program, 35-37

WAVES, enlisted training, 47; officer procurement, 34; officer training, description, 34, prediction of success, 186-190, 212
Women's Reserve (see WAVES)
Winchman and Hatchman Selection Test, 123

Z-transformation technique, 393